

< p > < name type=org > The Ministry of Truth < /name >, &mdash; < p > < name type=org lang=TECH > 

T	E	X	T	J	U	S	P	E	E	C	H	J
A	N	D	J	L	U	A	N	G	U	A	G	E
T	E	C	H	N	O	L	O	G	Y	J	J	

 < /name >, in < name type=org lang=TECH > 

T	E	C	H	N	O	L	O	G	Y	J	J
---	---	---	---	---	---	---	---	---	---	---	---

 < /name > target=N1 < rend=underline > &mdash; was startlingly different from any other object in sight. It was an enormous pyramidal structure of glittering white concrete,

## SYNTAX AND SEMANTICS OF PREPOSITIONS

soaring into the air. From where < name type=person > Winston < /name > stood it was just possible to read, picked out on its white face in elegant lettering, the three slogans of the < name type=person > Party < /name >: < q rend="centered caps" type=slogan > War is peace < /q > < q rend="centered caps" type=slogan > Freedom is slavery < /q > < q rend="centered caps" type=slogan > Ignorance is strength. < /q >

< /p >

< note place=foot id=N1 >

< name > Newspeak < /name > was the official

language of < name type=place > Oceania < /name >.

For an account of its structure and etymology see Appendix. < /note >



Springer

## Syntax and Semantics of Prepositions

# Text, Speech and Language Technology

---

## VOLUME 29

---

### *Series Editors*

Nancy Ide, *Vassar College, New York*  
Jean Véronis, *Université de Provence and CNRS, France*

### *Editorial Board*

Harald Baayen, *Max Planck Institute for Psycholinguistics, The Netherlands*  
Kenneth W. Church, *AT & T Bell Labs, New Jersey, USA*  
Judith Klavans, *Columbia University, New York, USA*  
David T. Barnard, *University of Regina, Canada*  
Dan Tufis, *Romanian Academy of Sciences, Romania*  
Joaquim Llisterri, *Universitat Autònoma de Barcelona, Spain*  
Stig Johansson, *University of Oslo, Norway*  
Joseph Mariani, *LIMSI-CNRS, France*

*The titles published in this series are listed on [www.springeronline.com](http://www.springeronline.com).*

# Syntax and Semantics of Prepositions

Edited by

Patrick Saint-Dizier

*CNRS, Toulouse, France*

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-3899-2 (PB)  
ISBN-13 978-1-4020-3899-0 (PB)  
ISBN-10 1-4020-3849-6 (HB)  
ISBN-13 978-1-4020-3849-5 (HB)  
ISBN-10 1-4020-3873-9 (e-book)  
ISBN-13 978-1-4020-3873-0 (e-book)

---

Published by Springer,  
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

[www.springeronline.com](http://www.springeronline.com)

*Printed on acid-free paper*

All Rights Reserved

© 2006 Springer

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed in the Netherlands

# Contents

Preface	xi
Acknowledgments	xiv
1	
Introduction to the Syntax and Semantics of Prepositions	1
<i>Patrick Saint-Dizier</i>	
1.    The class of prepositions	1
2.    About the syntax of prepositions	2
3.    Polysemy and sense restrictions	10
4.    Representing the semantics of prepositions	12
5.    Prepositions and multilingualism	20
6.    Overview of the book structure	20
References	23
2	
Preposition Contractions in Quebec French	27
<i>Luc V. Baronian</i>	
1.    Introduction	27
2.    Facts	28
3.    Two possible analyses	31
4.    External evidence for Analysis 2	32
5.    Core linguistic arguments for Analysis 2	33
6.    Conclusion: consequences of the analysis	36
Acknowledgments	40
References	41
3	
The A's and BE's of English Prepositions	43
<i>Andrew McMichael</i>	
1.    Some definitions	43
2.    Corpus data	44
3.    The general formative principle	45
4.    Adverb formation	47
5.    The A's: a more complicated origin	47

6.	Simplex and compound prepositions: a classification	48
7.	The prefixed Gp as a cognitive functional marker	49
8.	Cognitive schemata of grammaticalising prepositions: an alternative categorisation	49
9.	Extensions of the Pattern	52
10.	Language Typology	52
11.	Conclusion	54
	References	55
 4		
	Typological Tendencies and Universal Grammar in the Acquisition of Adpositions	57
<i>David Stringer</i>		
1.	Introduction	57
2.	A monkey, a parrot and a banana	58
3.	Lexical variation and single syntax	62
4.	Conclusion	66
	References	67
 5		
	Multilingual inventory of interpretations for postpositions and prepositions	69
<i>Mikel Lersundi and Eneko Agirre</i>		
1.	Previous work	70
2.	Method to obtain the inventory and the multilingual table	72
3.	Case study with the Basque instrumental postposition	74
4.	Overall results	77
5.	Remaining problems.	77
6.	Conclusions and future work	79
	Acknowledgments	81
	References	81
 6		
	German prepositions and their kin	83
<i>Martin Volk</i>		
1.	Introduction	83
2.	German prepositions	84
3.	Conclusions	93
	References	94
	Appendix: Prepositions	96
	Appendix: Contracted Prepositions	98
	Appendix: Pronominal Adverbs	98
	Appendix: Reciprocal Pronouns	99

7	
Directionality Selection	101
<i>Marcus Kracht</i>	
1. Introduction	101
2. Modes	102
3. One Word — Three Meanings	104
4. Selection	106
5. Significance for Interpretation	108
6. Predicting Selectional Properties	110
7. Mode Heads: Evidence from Mari	112
8. Conclusion	113
References	113
8	
Verb-Particle Constructions in the World Wide Web	115
<i>Aline Villavicencio</i>	
1. Introduction	115
2. VPCs in a Nutshell	118
3. VPCs and Dictionaries	118
4. VPCs and Corpora	121
5. VPCs in the Web	122
6. Conclusions	127
Acknowledgments	128
References	128
9	
Prepositional Arguments in a Multilingual Context	131
<i>Valia Kordoni</i>	
1. Introduction	131
2. The Data	132
3. Previous Accounts in HPSG	136
4. Indirect Prepositional Arguments: The Analysis	137
5. Conclusion	142
Acknowledgments	143
References	143
10	
The syntax of French <i>à</i> and <i>de</i> : an HPSG analysis	147
<i>Anne Abeillé, Olivier Bonami, Danièle Godard and Jesse Tseng</i>	
1. Introduction	147
2. Syntactic properties	150
3. Proposed HPSG analysis	154
4. Concluding remarks	159
References	161

11	
In Search of a Systematic Treatment of Determinerless PPs	163
<i>Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger and Ivan A. Sag</i>	
1. Introduction	163
2. The Syntax of Determinerless PPs	165
3. The Semantics of Determinerless PPs	169
4. Analysis	172
5. Conclusion	176
Acknowledgments	177
References	177
12	
Combinatorial Aspects of Collocational Prepositional Phrases	181
<i>Beata Trawiński, Manfred Sailer and Jan-Philipp Soehn</i>	
Introduction	182
1. Syntactic Aspects	182
2. Semantic Aspects	185
3. Irregular Combinations	190
4. Summary	193
Acknowledgments	194
References	195
13	
Distributional Similarity and Preposition Semantics	197
<i>Timothy Baldwin</i>	
1. Introduction	197
2. Calculating inter-preposition similarity	199
3. Gold standard sources of inter-preposition similarity	200
4. Evaluation	203
5. Related research	206
6. Conclusion	207
Acknowledgments	207
References	208
14	
A Computational Model of the Referential Semantics of Projective Prepositions	211
<i>John Kelleher and Josef van Genabith</i>	
1. Introduction	212
2. The Challenges	212
3. Previous Computational Work	215
4. The LIVE Model	216
5. Conclusions	226
References	227

15	
Ontology-Based Semantics for Prepositions	229
<i>Per Anker Jensen and Jörgen Fischer Nilsson</i>	
1. Introduction	229
2. Formal ontologies	231
3. The relation between lexicon and ontology	232
4. Formal meaning ascription	234
5. Generative ontologies with feature structures	235
6. Ontological affinities and generative ontologies	236
7. Compositional ontological semantics for nominals	237
8. Prepositions and semantic roles in Danish	238
9. Identifying paraphrases	241
10. Conclusion	242
Acknowledgments	243
References	243
16	
Analysis and Interpretation of the Japanese Postposition <i>no</i>	245
<i>Ryusuke Kikuchi and Hidetosi Sirai</i>	
1. Introduction	246
2. Syntactic Analysis of <i>No</i>	247
3. Semantic Analysis of <i>No</i>	248
4. Framework of Context-Dependent Interpretation — SDRT	249
5. Case Study	250
6. Discussion	256
7. Conclusion	258
References	259
17	
What do the notions of instrumentality and of manner have in common?	263
<i>Alda Mari</i>	
1. Aim and methodology	263
2. Analysis of <i>avec</i> -instrument and <i>avec</i> -manner	265
3. The model: properties and constraints	274
4. A model for <i>avec</i> -instrument and manner	279
5. Conclusion	283
References	286
18	
A Conceptual Semantics for Prepositions denoting Instrumentality	289
<i>Alda Mari, Patrick Saint-Dizier</i>	
1. An analysis of the primitive notion of instrumentality and its lexicalizations	289
2. Analysing the notion of instrumentality via its lexicalizations	291

3.	The logical model	295
4.	LCS representation of preposition senses and instances	298
5.	Conclusion	302
Acknowledgments		304
References		304
19		
Prepositions in Cooperative Question-Answering Systems: a Preliminary Analysis		307
<i>Farah Benamara and Véronique Moriceau</i>		
1.	Introduction	307
2.	Preposition use in Question Answering : the WEBCOOP system	309
3.	Semantic Representation and Interpretation of Localization Prepositions in WEBCOOP	311
4.	Reasoning with Localization Prepositions	317
5.	Generating Prepositions and PPs	323
6.	Conclusion	328
References		329
Index		331

# Preface

A great deal of attention has been devoted in the past ten years in the linguistic and computational linguistics communities to the syntax and the semantics of nouns, verbs and also, but to a lesser extent, to adjectives. Related phenomena such as quantification or tense and aspect have motivated a number of in-depth studies and projects. In contrast, prepositions have received less attention. The reasons are quite clear: prepositions are highly polysemic, possibly more so than adjectives, and linguistic realizations are extremely difficult to predict, not to mention the difficulty of identifying cross-linguistic regularities. Furthermore, a number of languages do not use prepositions or postpositions (or make a limited use of them) and prefer other linguistic forms such as morphological marks, e.g. case marks.

Let us mention, however, projects devoted to prepositions expressing space, time and movement in artificial intelligence and in natural language processing, and also the development of formalisms and heuristics to handle prepositional phrase attachment ambiguities. Prepositions are also present in subcategorization frames of predicative lexical items, but often in an informal and coarse-grained way. Let us also mention the large number of studies in psycholinguistics and in ethnolinguistics around specific preposition senses. Finally, prepositions seem to reach a very deep level in the cognitive-semantic structure of the brain: cognitive grammar developers often use prepositions in their metalanguage, in order to express very primitive notions. An important and difficult question to address, is whether these notions are really primitive or can be decomposed and lexically analysed.

In argument structure, prepositions often play the crucial role of a mediator between the verb's expectations and the semantics of the nominal argument. The verb-preposition-noun semantic interactions are very subtle, but totally crucial for the development of an accurate semantics of the proposition. Languages like English have verbal compounds that integrate prepositions (compositionally or as collocations) while others, like Romance languages or Hindi either incorporate the preposition or include it in the prepositional phrase. All these configurations are semantically as well as syntactically of much interest.

Prepositions turn out to be a very useful category in language, it does not just play the role of a grammatical marker. Prepositions are essential in a number of applications such as indexing and knowledge extraction since they convey basic meanings of much interest like instruments, means, comparisons, amounts, approximations, localizations, etc. They must necessarily be taken into account—and rendered accurately—for effective machine translation and lexical choice in language generation.

Prepositions are also closely related to semantic structures such as thematic roles, semantic templates or frames, and subcategorization frames. From a linguistic perspective, several investigations have been carried out on quite diverse languages, emphasizing e.g., monolingual and cross-linguistic contrasts or the role of prepositions in syntactic alternations. These observations cover in general a small group of closely related prepositions. The semantic characterization of prepositions has also motivated the emergence of a few dedicated logical frameworks and reasoning procedures.

This book emerges from a workshop on the syntax and semantics of prepositions, organized in Toulouse in September 2003. The aim of this workshop was to bring together linguists, NLP researchers and practitioners, and AI people in order to define a common ground, to advance the state-of-the-art, to identify the primary issues and bottlenecks, and to promote future collaborations. The main topics were:

- The syntax of prepositions: formal or descriptive syntax, prepositions in alternations, principles in the syntax of PPs, syntactic and semantic restrictions. General syntactic-semantic principles. Postpositions or other equivalent markers (e.g. case).
- Descriptions: Potential WordNet / EuroWordNet descriptions of preposition uses, productive uses versus collocations, multi-lingual descriptions: mismatches, incorporation, divergences. Prepositions and thematic roles, prepositions in semantic frameworks (e.g. Framenet.).
- Cognitive or logic-based formalisms for the description of the semantics of prepositions, in isolation, and in composition / confrontation with the verb and the NP. Compositional semantics. Logical and reasoning aspects.
- Cognitive or logic-based formalisms for the description of the semantics of prepositions, in isolation, and in composition/confrontation with the verb and the NP. Compositional semantics. Logical and reasoning aspects.
- The role of prepositions in applications, in particular: in machine translation, in information extraction, and in lexicalization in language generation.

- Corpus-based studies that support or challenge any of the approaches described above.
- Lexical knowledge bases and prepositions. Prepositions in AI, KR and in reasoning procedures.

The Program Committee, that we warmly thank, was the following :

Nicholas Asher (Austin)  
Pushpak Bhattacharyya (IIT, Mumbai)  
Harry Bunt (Tilburg)  
Nicoletta Calzolari (Pisa)  
Bonnie Dorr (Maryland)  
Christiane Fellbaum (Princeton)  
Claire Gardent (CNRS, Nancy)  
Betsy Klipple (Upenn)  
Alda Mari (ENST, Paris)  
Palmira Marrappa (Lisboa)  
Martha Palmer (Upenn)  
Patrick Saint-Dizier (IRIT, Toulouse)  
Gloria Vazquez (Lerida)  
Laure Vieu (IRIT, Toulouse)

PATRICK SAINT-DIZIER

## Acknowledgments

We thank the programme committee, cited in the preface, for their useful contributions, and two anonymous reviewers who made useful comments on the whole book.

We also thank our 2 sponsors : *Université Paul Sabatier* and the *GDR-CNRS Sémantique* and the Organizing Committee members: Dr. Farah Benamara, Véronique Moriceau and Sara Mendès.

# Chapter 1

## INTRODUCTION TO THE SYNTAX AND SEMANTICS OF PREPOSITIONS

Patrick Saint-Dizier

*IRIT-CNRS*

*118 Route de Narbonne 31062 Toulouse, France*

[stdizier@irit.fr](mailto:stdizier@irit.fr)

**Abstract** This first chapter presents basic issues related to preposition syntax and semantics. It introduces different ways to view the syntax of prepositions: relational, functional and lexical. It also shows the high degree of polysemy of a number of prepositions and develops some directions to deal with preposition semantics, in particular designed for natural language processing systems, based on the Lexical Conceptual Structure and underspecification.

**Keywords:** syntax and semantics of prepositions.

### 1. The class of prepositions

Prepositions do not exist in all languages. While some languages, such as Indian languages (Hindi, Telugu, Tamil, etc.), have postpositions rather than prepositions, but this may be viewed as a rather minor distinction, other languages do not have prepositions but e.g. morphological marks such as cases, which play an equivalent role. Prepositions do not form a strict closed class of elements, as sometimes hastily presented by grammarians. Most languages with prepositions have a rather limited set of single word prepositions, in general between 40 and 120, although there are divergences among grammarians on the exact nature and definition of a preposition. In addition, there is quite large number of prepositional compounds, i.e. structures that play the role of prepositions, that include nouns (*sur le côté de*, *on the left of*, *al lado de* (Fr., Eng., Sp.)), adjectives (*proche de*, *close to*) or gerundives (*se rapportant à*, *with respect to*). Finally, preposition uses are very different from one language

to another, even within closely related languages in a linguistic family, with often a large number of idiosyncratic constructions: *dream about*, *rêver de* (litt. dream ‘of’ in French), *soñar con* (litt. dream ‘with’ in Spanish and in Portuguese). Not surprisingly, a number of prepositions are highly polysemic, almost comparable to the most polysemous adjectives like *good*.

The fact that cases or other morphemes or affixes are used in some languages instead of prepositions indicates that prepositions have specific relations with other types of linguistic mechanisms. Let us now investigate the different roles played by prepositions from a syntactic and semantic point of view.

Prepositions can first be viewed as a **functional category** in syntax: they are heads of prepositional phrases. The preposition then hierarchically dominates the noun phrase. Prepositions can also be viewed as a **semantic relation** between a structure that precedes it (e.g. a verb) and another one that follows it (e.g. an NP). This relation can be represented as a conceptual relation, as shall be seen below. Finally, prepositions can be viewed as a **lexical category** that imposes both a categorial (structure level) and a semantic selection (semantic restriction level). Similarly to the other predicative categories, prepositions have type restrictions on their arguments, they assign thematic roles, and they have a semantic content, possibly underspecified. The only difference with the other open-class categories like nouns, verbs or adjectives is that they do not have any morphology. These considerations show the central role played by prepositions in the proposition and their fundamental predicative and relational nature.

In the following sections we present some aspects of the syntax and the semantics of prepositions. These are basic notions meant for the reader unfamiliar with prepositions. A number of these notions are further developed in the following chapters for particular classes of prepositions, or for particular languages.

## 2. About the syntax of prepositions

There are only about 50 prepositions in English (for other languages there is not always a consensus on what a preposition is, e.g. vs. prepositional compounds). Here is a fairly complete list: *aboard, about, above, across, after, against, along, amid, among, anti, around, as, at, before, behind, below, beneath, beside, besides, between, beyond, by, despite, down, during, except, excepting, excluding, following, for, from, in, inside, into, like, near, of, off, on, onto, opposite, outside, over, past, per, plus, round, save, since, than, through, to, toward, towards, under, underneath, unlike, until, up, upon, versus, via, with, within, without*.

In this section, we investigate the different facets of the syntax of prepositions: phrasal constructions with prepositions, prepositions as relations, prepositions as thematic role assignators and prepositions in alternations.

## 2.1 Preposition distribution in English and French

Before going into the details of the syntax of prepositions, let us say a few words about preposition distribution, illustrated here on English and French.

The WFWSE web site indicates that English prepositions (on the lexeme basis) are distributed as follows in ordinary, everyday English. Among the 30 most frequent words in English, there are 9 prepositions:

Fig. 1a - Preposition uses in English	
preposition	rank
OF	2
IN	5
TO	8
FOR	11
WITH	13
ON	16
BY	18
AT	20
FROM	29

Rank indicates here the usage rank of the term all words considered. For example, of is the second most frequently used word in English.

For French, we have collected 14656 preposition usages from various corpora, their relative occurrence frequencies, within the set of all French prepositions, are distributed as follows:

Fig. 1b - Preposition frequencies in French		
preposition	occurrences	frequency (%)
DE, DES, D', DU (of)	8338	57
A, AU, AUX (at, to)	1649	11.2
EN (of)	856	5.8
POUR (for)	719	4.9
SUR (on)	704	4.8
DANS (in)	462	3.1
PAR (by)	413	2.8
AVEC (with)	280	1.9
ENTRE (between)	85	0.57
VERS (towards)	67	0.46
SOUS (under)	66	0.45
CONTRE (against)	62	0.44

The other prepositions (e.g. *east of, above, along*) occur less than 50 times, in general less than 10 times. If we do not take into account DE and A and

their morphological variants, frequencies need to be multiplied by 3.14 (no relation with the number  $\pi$ , though). The observation is that 16 prepositions occur more than 1%. They are not necessarily the most polysemic ones (e.g. *entre* (between) is not very polysemic).

## 2.2 Phrasal verbs

Phrasal verbs, also called prepositional verbs are verb + preposition constructions. These constructions may range from purely idiosyncratic forms (*boil down*) to compositional ones (*switch on*, *run into*). In the first example, the preposition has an intransitive use, whereas in the second it has a transitive use, where the NP is missing, possibly elliptical, but can be reconstructed, e.g. via inference. Other cases include, for example, making explicit an information which would by-default be incorporated. For example, in *climb down*, the preposition ‘down’ is made explicit because the by-default incorporated preposition is *up*.

Non compositional phrasal verbs are common, for example, in English and German (e.g. *ab-stammen*, *auf-nehmen*); they are less frequent in Romance languages, which mainly allow transitive uses (*Il est tombé dessus*, he fell on). Non compositional verb + preposition compounds, also termed verb particle construction (see e.g. Villavicencio, this volume), are often viewed as a lexical unit *per se*, which can subcategorise for a PP or an NP, as in:

(1a) *John switched on the light*

where ‘switched on’ subcategorises for an NP. In the case of a phrasal verb where the association verb + preposition is compositional, a useful (but not systematic) test is that the order of the preposition and the NP can be switched around:

(1b) *John switched the light on.*

Which is neither possible with idiosyncratic forms:

\* *This talk boils to very few concrete propositions down.*

nor with prepositions in regular PPs:

\* *Mary is waiting John for.*

In most computational linguistics approaches, phrasal verbs are considered as separate lexical units: their subcategorization frame(s), possible alternations and other syntactic properties are described in dedicated lexical entries. It is indeed very difficult to generalize lexical behavior for a given preposition and all the verbs with which it can be combined.

## 2.3 Prepositions as relations

In general, prepositions introduce a relation between two entities or sets of entities. The first entity is often a kind of external argument while the second

one is headed by the preposition. In *Mary goes to school*, to has two arguments: Mary (external) and school: to(Mary, school). Mary is an argument shared with the verb go.

Prepositions select in general NPs but also sometimes propositions. In some cases, NPs or propositions can be omitted, they are however implicit and can be inferred from the context. Prepositions, as shall be seen below, have their own selectional restrictions. In a VP construction (V PP), the selectional restrictions imposed by the verb on its indirect object (PP) must in some way coincide with the type of the PP (e.g. direction, instrument) and with the type of the NP within the PP. Consider a simple illustration:

(2) *to run to school*

In (2), *run* requires a path, probably underspecified w.r.t. the area in which it occurs. This requirement is met by the preposition *to*. In turn, *to* expects an NP of type: closed, well-delimited, possibly large, space. *School* meets these requirements.

Prepositions such as *around*, *out*, *in*, *away* can be used with empty objects: *go away*, *stroll around*, even if the object is in fact implicit, possibly vague. Prepositions such as *in*, *into*, *without* select an NP complement: *in the room*, *without sugar*, while prepositions such as *out*, *from* can select NP or PP complements: *from under the table*, *out in the streets*. Finally, prepositions such as *between* select a plural NP: *between John and Mary*, *between my 5 best friends*. Finally, a few authors tend also to consider that prepositions such as *from* or *down* select two NPs, as in *from A to B*, *down A to C*. We think this analysis is not correct because e.g. *from* only selects A. The expressions *from A to B* must be analysed as a compound of type *trajectory* where *from* and *to* play an equivalent role.

Besides the NP or PP it selects, a preposition has a kind of ‘external’ argument, possibly shared with another predicate, which is the first element of the relation:

(a book) on (the table).

(3) (Mary) entered into (the opera house).

A more complex case includes two intertwined relations:

(4a) (Max) steals sweets from (behind the counter). and

(4b) Max steals (sweets) from behind (the counter).

In (4a), *steals* indeed expects e.g. a kind of trajectory describing the path followed by the stolen object, whereas in (4b) *sweets* are analysed as being in a fixed position, specified by the preposition *behind*.

Prepositions can, similarly to verbs, be associated with a subcategorization frame where the first element of the frame is in general shared with another predicate, in general a verb. We can then have in a lexical entry the following description for the accompaniment sense of *with*:

with : [NP, NP]

Selectional restrictions can be added to that frame, on each argument position, with the same well-known accuracy problems as for verbs. In particular, a number of preposition senses (at least half of them) can be subject to several forms of metaphors (Moriceau et al. 03).

## 2.4 Prepositions and thematic roles

Thematic roles are abstract labels that characterize the semantic relations between predicates and their arguments. Each argument of a predicate is marked with a thematic role, which indicates, in a very general way, the ‘semantic’ role played by the argument with respect to the predicate. From this point of view, thematic roles can be considered as a first level of semantic representation, of much interest, for example, in knowledge extraction, where it may not be possible to go much deeper in the semantics, due to the size of the explored documents. Thematic roles have been subject to many controversies, and there is still little agreement on their nature, definition, and role in linguistic theories (Gruber 67), (Jackendoff 87), (Rappaport and Levin 88), (Roca 92), (Ravin 90).

Here is a partial list of roles, which is however generally agreed upon:

- **agent**: the entity who intentionally initiates, makes or originates the action described by the predicate,
- **patient**: the entity that undergoes the action described by the predicate, it is often an animate entity,
- **theme**: the entity moved (in a very general sense) as a consequence of the action expressed by the predicate, it is often a non-animate entity,
- **experiencer**: the entity that experiences some psychological state resulting from the predicate,
- **goal, source, location**: are roles related to spatial, temporal or abstract fields, expressing respectively the goal, the source or the position of a temporal, spatial or abstract entity.

Thematic roles are postulated by a number of authors to be universal, non-ambiguous, and to cover the whole spectrum of the predicate-argument relationships. This is certainly somewhat optimistic. Thematic roles are essentially assigned to NPs, by verbs, prepositions and VPs via predication. Their uses and meanings may be either direct or metaphorical. For example, meteorological forces are often metaphorically assimilated to agents: *the wind broke the window*.

Prepositions can be associated with a thematic grid, which contains in general one role per argument position, but multiple assignments are also possible

when two or more roles are relevant, as for verbs. For example, the following prepositions have the following grids:

*on*: [theme, location].

*between*: [theme, location].

*towards*: [theme  $\vee$  agent, goal].

In general, a preposition assigns a thematic role to its ‘object’ argument, i.e. the argument in the scope of the PP it heads. Therefore, *towards* assigns the role goal to its object NP. A preposition being a relation, it is also necessary to take into account another argument, the first argument of the relation, a kind of ‘external’ argument, that the preposition shares in general with the verb of the proposition (or another type of predicate). This latter argument gets thematic roles from at least two sources, which must, obviously, be compatible.

Thematic roles can be defined a priori as by-default roles, which are assigned in sentences when there is no contradiction. However, in a number of situations, they can be revised, in particular in sense extensions, for example goal can become location. This is typical in systems with more refined thematic role typologies (Boguraev 79), (Dowty 89, 91), (Saint-Dizier 99). For example, in:

*The arrow moves towards the target*: the external argument is a theme,  
*John runs towards the restaurant*: John is an agent.

This problem can be solved by leaving the first role underspecified or by listing all the possibilities in the lexical entry of *towards*. In general, the role(s) mentioned a priori is(are) the most prototypical.

## 2.5 Prepositions and PP attachment ambiguities

Since the very beginning of language processing techniques, the management of PP-attachment ambiguities has been a real challenge, for which no fully satisfactory solution has ever been proposed. One of the reasons is that resolving such ambiguities often requires non trivial contextual inferences, similarly to e.g. reference resolution (remember the well-known example *I saw a man with a telescope in the park*).

However, the development of large ontologies, used to type in a relatively accurate way predicate arguments and the introduction of heuristics or preferences (based e.g. on statistical analysis and learning techniques) allowed significant progress in this area. If attachment cannot be resolved at parse time, a common approach is to produce a syntactic (or semantic) representation that allows the representation of the ambiguity (e.g. by means of multiple links in syntactic trees which become locally graphs). The ambiguity may then be resolved during the interpretation.

## 2.6 Prepositions in syntactic alternations

In her book, Beth Levin (Levin 93) shows, for a large set of English verbs (about 3200), the correlations between the semantics of verbs and their syntactic behavior. More precisely, she shows that some facets of the semantics of verbs have strong correlations with the syntactic behavior of these verbs and with the interpretation of their arguments. This very important work emerged from the synthesis of specific investigations on particular sets of verbs (e.g. movement verbs), on specific syntactic behaviors and on various types of information extracted from corpora. Other authors have studied in detail the semantics conveyed by alternations, e.g. (Pinker 89) and the links between them (Goldberg 94).

**2.6.1 The alternation system.** An alternation, roughly speaking, describes a change in the realization of the argument structure of a verb. The scope of an alternation is the proposition. Modifiers are considered in some cases, but the main structures considered are the arguments, including prepositions, and the verb. Arguments may be deleted or ‘moved’, NPs may become PPs or vice-versa, and some PPs may be introduced by a new preposition. Alternations may also be restricted by means of constraints on their arguments.

Beth Levin has defined 79 alternations for English. They basically describe ‘transformations’ from a ‘basic’ form. However, these alternations have *a priori* little to do with the assumptions of Government and Binding theory and Movement theory, in spite of some similarities. The form assumed to be basic usually corresponds to the direct realization of the argument structure, although this point of view may clearly be subject to debate. Here are now a few types of alternations, among the most common ones. References about works establishing these relations can be found in (Levin 93).

The *Transitivity alternations* introduce a change in the verb’s transitivity. In a number of these alternations the subject NP is deleted and one of the objects becomes the subject, which must be realized in English. The *Middle alternation* is typical of this change:

*John cuts the cake* → *The cake cuts easily.*

As can be noticed, it is often necessary to add an adverb to make the sentence acceptable. The *Causative/inchoative alternation* (Levin 93) concerns a different set of verbs:

*Edith broke the window* → *The window broke.*

Verbs undergoing this alternation can roughly be characterized as verbs of change of state or position.

Under the transitivity alternations fall also alternations where an object is unexpressed. This is the case of the *Unexpressed object alternation* where the object1 is not realized. A number of verbs undergo this alternation. In most

cases, the ‘typical’ object is somewhat ‘implicit’ or ‘incorporated’ into the verb, or deductible from the subject and the verb. This is the case, e.g., for the *Characteristic property of agent alternation*:

*This dog bites people* → *This dog bites*.

**2.6.2 Alternations involving prepositions.** An interesting alternation, with a heavy semantic impact, is the *conative alternation* that changes the object NP into a PP introduced in English by *at* (*sur* in French), as in:

*Edith cuts the bread* → *Edith cuts at the bread*.

A second set of alternations deals with changes within the arguments of the VP. One of the most popular alternations is certainly the *Dative alternation* which concerns verbs of giving, of future having, of transfer, etc., as in:

*Edith hands the baby a toy* ↔ *Edith hands a toy to the baby*

(we use the symbol ↔ for some examples when we feel that both forms have equal status, i.e. one or the other could be considered as basic). The same phenomenon occurs for the *Benefactive alternation*:

*I carve a toy for the baby* ↔ *I carve the baby a toy*.

The *Spray/Load alternation* involves the permutation of the arguments in the VP and the preposition alternation on ↔ with:

*to spray paint on the wall* ↔ *to spray the wall with paint*.

English is particularly rich in this type of phenomenon. Let us note also the *Material / product alternation*:

*Martha carves a toy out of a piece of wood* ↔ *Martha carves a piece of wood into a toy*,

and the *With / Against alternation*:

*to hit a stick against the fence* ↔ *to hit the fence with a stick*.

**2.6.3 The location alternations.** The location alternations, a family of alternations which involve a permutation of object1 and object2 and a preposition change, are also of much interest. The participation to certain of these alternations allows one to predict the type of motion and the nature of the end state. Verbs which focus only either on the motion (e.g. pour) or on the resulting state (e.g. fill) do not alternate. Verbs that alternate constrain in some manner both motion and end state. Let us now specify in more depth these constraints, since in fact quite a few verbs do alternate.

For example, let us consider the *into/with* alternation. (Pinker 89) differentiates among verbs which more naturally accept the *into* form as their basic form and which alternate with a *with* form. Their general form is:

*Verb NP(+theme) onto NP(+destination)*, and they alternate in:

*Verb NP(+destination) with NP(+theme)*.

*Load hay into the wagon / Load the wagon with hay.*

Other verbs more naturally take the location/container as object (e.g. stuff), their basic form is more naturally:

*Verb NP(location) with NP(+theme), and alternate in:*

*Verb NP(+theme) onto NP(+destination).*

*stuff the truck with hay / stuff hay onto the truck.*

Verbs which undergo the ‘into/onto’ alternation, have one of the following properties: simultaneous forceful contact and motion of a mass against a surface (brush, spread, ...), vertical arrangement on a horizontal surface (heap, pile, stack), force is imparted to a mass, causing ballistic motion along a certain trajectory (inject, spray, spatter), etc. Those which do not alternate have, for example, one of the following properties: a mass is enabled to move via gravity (spill, drip, spill), a flexible object extended in one direction is put around another object (coil, spin, twist, wind), a mass is expelled from inside an entity (emit, expectorate, vomit). As can be seen here, the properties at stake are very precise and their identification is not trivial, especially for verbs which can be used in a variety of utterances, with some slight meaning variations.

In general, alternations are described at a global level, and each verb is associated with the alternations it undergoes. Preposition changes are thus specified at this level. *A priori*, there is no specific information encoded in the preposition lexical entries.

### 3. Polysemy and sense restrictions

In this section we briefly evoke the problem of polysemy, crucial for prepositions, and the difficulty of characterizing sense boundaries, in particular by means of selectional restrictions. Representation issues are presented in the next section, after these preliminaries.

#### 3.1 Prepositions and polysemy

It is well known that prepositions are highly polysemic and enter into a large number of metonymies and metaphors. However, some prepositions have very restricted uses such as *west of*, *in order to*, *during*, *in spite of*, *in favor of*, *except*, *thanks to*, *concerning*, *via*, etc. We believe that it should be possible to identify a reasonable number of ‘kernel’ senses for each preposition, that accommodate several forms of variations.

The identification of a preposition sense needs to be based on the observation of groups of usages. Two criteria must be taken into account: (a) the nature and the stability within a certain semantic domain related to the head noun type of the PP controlled by the preposition, that confirms the ontological basis of the sense and, concomitantly, (b) the restrictions required by the verb on the

nature of the PP, if it is an argument. Dictionary definitions and multilingual considerations may also help. Pragmatic factors may also interfere, but this is much more ad'hoc.

Although prepositions have some idiosyncratic usages (probably much less in French than in English), most senses are relatively generic and it should be possible to characterize them using relatively consensual and high-level ontology labels.

Let us consider the case of the French preposition *par*. We have identified six senses which can be identified and characterized as shown below. These senses occur in very diverse ontological domains while being all approximately at the same level of abstraction:

- proportion or distribution: *il gagne 1500 Euros par mois* (he earns 1500 Euros per month),
- causality: as in passives but also e.g. in *par mauvais temps, je ne ne sors pas* (by (=in) bad weather I don't go out),
- origin: *je le sais par des amis* (I know it from friends),
- via: *je passe par ce chemin* (I go via (=by) this path),
- tool or means: *je voyage par le train* (I travel by train),
- approximation of a value: *nous marchons par 3500m d'altitude* (we hike at an altitude of 3500m).

An important point is that uses of *par* do not necessarily cover all the conceptual field which could ‘naturally’ be associated with each sense. For example, the expression of the idea of approximation using *par* is rather restricted to localization, speed or movement, it does not include e.g. amounts. One of the tasks is then to characterize, for each sense, what the subset of the conceptual field is. This is done by two means: (1) by a semantic characterization of the NP dominated by the preposition and (2) by the analysis of the restrictions imposed by the verb of the clause on the PP, or, conversely, by the type or the family of the verb (e.g. possession, communication, as in WordNet) the preposition can be combined with, for that particular sense.

Let us now examine the basic restrictions for three senses of *par*. The ‘VIA’ sense is basically subcategorized by movement verbs; it is a path, subcategorizing for a noun of type ‘way’ or ‘route’ or, by a kind of metonymic extension, any object which can define a trajectory, e.g. an aperture (by the window). It has numerous metaphors in the psychological and epistemic domains (e.g. *Il passe par des moments difficiles* (He experiences difficult moments)).

The ‘ORIGIN’ (or ‘SOURCE’) sense is more narrow, it is essentially used in conjunction with communication or epistemic verbs, the argument is usually

of type place, and the head noun is of type ‘human’: *Il transite par Paris* (he commutes in Paris). We consider that nouns of type e.g. ‘object with an informational content’ or ‘human’ introduce a metonymic extension, as in, e.g. *par la radio / la presse / des amis* (I know the news from the radio / the newspapers / friends).

Finally, the ‘TOOLS or MEANS’ sense is used with verbs describing concrete actions (e.g. creation and movement verbs, if we refer to the verb class system of WordNet (Fellbaum 93)). In general it is an adjunct. It is typed as a means, and the head noun of the PP must be e.g. a tool, or, more generally, an object that allows the action to be realized, which may not necessarily be prototypical. This object could be found e.g. in the encyclopedic knowledge associated with the verb, or via a functional relation in a thesaurus. It has also numerous metaphoric extensions (e.g. *je traite ce phénomène par la logique temporelle* (I deal with this phenomena ‘by’ temporal logic)).

### 3.2 Some difficulties with selectional restrictions

However, there are many well-known difficulties inherent to the selectional restriction approach, where additional, non-trivial, world knowledge is required to make sense distinctions. Consider the usage:

‘Dans (in) followed by an NP of type location’ (e.g. to be in a drawer).

*Location* is obviously too general a restriction (\**to be in the shelf*). It is then necessary to enter into more complex descriptions, specifying that the location has a (salient) ‘inside’, that is not just a surface, etc. However, as far as only elementary spatial properties are concerned, this remains feasible.

More complex is the case of *boire dans un verre* (literally: drink in a glass). This example highlights the complex interactions between the verb and its PP. The preposition is part of the PP, not part of a verb complex form, this latter construction being quite unusual in French. The recipient is not neutral: while *verre, tasse, bol,...* (glass, cup, bowl) are acceptable arguments, *bouteille, robinet* (bottle, faucet) are not, probably because of their narrow neck, which prevents the drinker from having his mouth inside the recipient. This characterization becomes more complex and, probably, an interpretation for example in terms of Euclidean geometry could be necessary.

## 4. Representing the semantics of prepositions

A few general purpose classifications for prepositions have been proposed in the past. They tend, in most cases, to converge quite well. In this section, we survey two of them. The first was introduced in the eighties by (Boguraev and Spark Jones 87), while the latter serves as a basis for the PrepNet project (Saint-Dizier 05). Another classification, based on a lexicographic method-

ology, is presented at <http://www.clres.com/prepositions.html> in the Preposition Project (TPP).

In the remainder of this section, we focus on representation and expressivity issues. Basic underspecification techniques are introduced to show the verb-PP interactions in semantic composition.

## 4.1 A study of cases

Let us present first Boguraev and Sparck Jones classification. It is based on (Woods 79) which provides an extensive list of preposition uses in English. Their study was a purely investigative one, with the aim of characterizing sentence relations. They did not address the question of how the specific assignments, for each individual sentence, could be achieved. We give below the main elements of the list, whis is given in (Boguraev and Spark Jones 87), and accessible via the ACL digital library. Cases are not structured. For each of them, we give the prototypical English prepositions. They are given below in alphabetic order:

- Accompaniment (with),
- Activity (at),
- Abstract destination (to),
- After (after),
- Abstract location (in),
- Abstract source (from),
- Attribute (in, with),
- Before (before),
- Comparison (as),
- Destination (to),
- Direction (down, ...),
- Goal (for),
- Instrument (by, with),
- Location (at),
- Manner (with),
- Reason (because of),
- Source (from),
- Time location (at).

## 4.2 The PrepNet classification

Here is an organization of the different senses for prepositions as implemented in PrepNet, which is still in an early stage of developement (accessible at: [www.irit.fr/recherches/ILPL/prepnet.html](http://www.irit.fr/recherches/ILPL/prepnet.html)), with some frequent minor adjustments. Senses are called abstract notions, to dissociate them from linguistic realizations. The classification was initially elaborated from French (Cannesson

et al 02), but seems largely valid for most European languages. It also coincides to a large extent with other classifications, presented in some chapters of this volume.

Senses are organized on three levels:

- 1 a first level characterizes a **semantic family**, a level roughly comparable to thematic roles: localization, manner, quantity, accompaniment, etc.,
- 2 a second level accounts for the different **facets** of the semantic family, e.g. source, destination, via, fixed position for the localization family,
- 3 a third level characterizes, roughly speaking, the **modalities of a facet** when appropriate. For example, the facet *manner and attitudes* is decomposed into 3 modalities: *basic manner*, *manner by comparison* and *manner with a reference point*. Due to space limitations, this latter level will not be developed in this document.

It is also important to note that each preposition sense is considered from the point of view of its basic usage and as the source of numerous metaphors. For example, origin is basically spatial, but has numerous metaphorical transpositions into the temporal, psychological and epistemic domains, to cite just a few generic cases.

Here is the current PrepNet preposition classification, one or more examples follow to illustrate definitions, which cannot be given here in extenso due to space limitations:

■ **Localization** with subsenses:

- **source**,
- **destination**,
- **via / passage**,
- **fixed position**.

Destination may be decomposed into destination reached or not (possibly vague), but this is often contextual. From an ontological point of view, all of these senses can, a priori, apply to spatial, temporal or to more abstract arguments.

■ **Quantity** with subsenses:

- **numerical or referential quantity**,
- **frequency and iterativity**,
- **proportion or ratio**.

Quantity can be either precise (*temperature is 5 degrees above 0*) or vague. Frequency and iterativity, e.g.: *he comes several times per week*.

■ **Manner** with subsenses:

- **manners and attitudes**,

- **means (instrument or abstract),**
- **imitation or analogy.**

Imitation: *he walks like a robot; he behaves according to the law,*

■ **Accompaniement** with subsenses:

- **adjunction,**
- **simultaneity of events (co-events),**
- **inclusion,**
- **exclusion.**

Adjunction : *flat with terrace / steak with French fries / tea with milk,*

Exclusion: *they all came except Paul.*

■ **Choice and exchange** with subsenses:

- **exchange,**
- **choice or alternative,**
- **substitution.**

Substitution : *sign for your child, Choice: among all my friends, he is the funniest one.*

■ **Causality** with subsenses :

- **cause,**
- **goal or consequence,**
- **intention.**

Cause: *the rock fell under the action of frost.*

■ **Opposition** with two ontological distinctions: physical opposition and psychological or epistemic opposition. Opposition: *to act contrary to one's interests.*

■ **Ordering** with subsenses:

- **priority,**
- **subordination,**
- **hierarchy,**
- **ranking,**
- **degree of importance.**

Ranking : *at school, she is ahead of me.*

■ **Minor groups:**

- **About,**
- **in spite of,**
- **comparison.**

About: *a book concerning dinosaurs.*

Each of the facets described above is associated with a number of preposition lexicalizations. Here is a brief description of the Ordering family, with its 2 subsequent levels:

Fig. 2 - prepositions of the Ordering family		
facet	modality	preposition sense of
Priority	before	before / avant
	after	after / après
Subordination	under	under / sous
	above	on / sur
Hierarchy	under	before / derrière, avant
	above	front of, after / devant, après
Ranking	before	before, ahead of / devant
	after	after / derrière
Degree of importance	proximity comparison	near, close to / à côté de, auprès de, par rapport à, for, with respect to / pour, vis-à-vis de

## 4.3 Semantic representation and underspecification

Each preposition sense can be associated with a semantic representation, often largely underspecified. Let us consider in this chapter a simple illustration that shows some methodological elements and some basic difficulties, which will be deepened in the next chapters.

**4.3.1 Representing preposition senses.** Senses are described at two levels: (1) by means of a thematic grid characterizing the ‘standard’ function of each argument as presented in section 2 and, mainly (2) by means of a knowledge representation formalism, for example the Lexical Conceptual Structure (LCS) (Jackendoff 90, 97), which seems to be sufficiently expressive for that purpose. Compared to verbs, representing prepositions in LCS is rather straightforward and much more adequate. The difficulty is to elaborate a minimal, but sufficiently discriminatory set of primitives (55 in (Wierzbicka 92) system, 68 in (Cannesson et al. 02)). Y. Wilks introduces in (Wilks 77) the main arguments for and against the use of primitives, a long, recurring debate during the 70-80s.

A few principles guide this description: (1) the representation of generic senses (e.g. family level) subsumes the representation of their daughters, (2) different senses of a given preposition must receive substantially different semantic representations, (3) metaphoric uses are characterized in part by semantic field substitution in the LCS, not by a different representation with different primitives, and (4) the number of primitives representing prepositions must be as limited as possible. These primitives are lower in the LCS primitive hierarchy than e.g. the GO, CAUSE or BE primitives.

An important feature of the semantic representation of prepositions is the evaluation of an adequate level of genericity, that includes a number of variations related to the semantics of the preposition arguments. A possible solution consists in associating LCS representations with:

- a typed- $\lambda$ -calculus for the identification and characterization of underspecified fields and for semantic composition and
- logical devices to represent and constrain underspecification (e.g. defaults, constrained choices).

Let us first consider how primitives are elaborated on. To give a flavor of their descriptive level, here are a few of them, definitions in English being quite informal:

Fig. 3 - A few LCS primitives for prepositions	
primitive	short definition
ABOUT	concerning, theme of verb
ABOVE	fixed position above something, no contact
ON	same as ABOVE but with contact
AS	manner via imitation
AT	fixed, precise localization no notion of container
CLOSE-TO	in neighbourhood, no contact
EXCEPT	exclusion
DURING	expression of a duration
END	fixed loc. at end of
INSTEAD	substitution, replacement
PER	reference, for a frequency
AROUND	area around another area
AMONG	selection in a set
CO	accompaniment, co-events
NEXT-TO	immediate proximity, possible contact adjacency
THROUGH	movement via a narrow passage
VIA	movement via an unconstrained area

These primitives are directly preposition names in the LCS meta-language, but they are not necessarily used directly for the corresponding preposition. For example, two major senses of the preposition *avec* (with) (Mari 00) are:

- **accompaniment - simultaneity of events**, represented as:

$\lambda I [_{manner} CO_{+loc} ([_{thing} I])]$ ,

+loc indicates a physical accompaniment (*I go to the movies with Maria*), while +psy instead of +loc indicates a psychological accompaniment (*Maria investigated the problem with Joana*).

- **Manner - means - instrument**, represented as:

$\lambda I [_{manner} BY - MEANS - OF ([_{thing} I])]$

(*they opened the door with a knife*). This is, in fact, a generic representation for most prepositions introducing instruments (realized as: *à, à l'aide de, au moyen de, avec, par* (by means of, with, by, thanks to)).

Note that both senses are contrasted by different selectional restrictions on the NP, represented by the variable I.

More subtle is the representation of *contre* in French (approximately ‘against’, glosses in English are provided to facilitate reading), for which we give the comprehensive representation of its five senses:

- A first sense describes a **physical object positioned against another one** (in the hierarchy above: localization - fixed position - spatial):  
 $\lambda K [place \text{ } NEXT - TO_{+loc,c:+}([thing \text{ } K])]$   
 where NEXT-TO indicates a physical (+loc) proximity; contact is encoded by c:+ between two objects I and K, where I is against K. The analysis is that *contre* describes a position, not a path. It is important to note that the idea of movement, if any (as in: *push the chair against the wall*), comes from the verb, not from the preposition.
- *Contre* is also used to express **opposition**: *to swim against the current* or, metaphorically in the epistemic or psychological domains: *to argue against a theory / a practice*. The primitive OPPOSITE is used to capture the fundamental idea of antagonistic forces:  
 $\lambda K [place \text{ } OPPOSITE_{+locV+psV+epist,c:-,ta:+}([thing \text{ } K])]$ .  
 In that case, the physical contact is not relevant (c:-), while the agonist / antagonist force is present (noted ta:+, (Jackendoff 90), slightly simplified here).
- *Contre* can also be used to express notions like **providing a certain protection or defense** in the hierarchy ‘causality - goal’: *medecine for cough*. It is represented as follows:  
 $\lambda X [eventVstate \text{ } FOR([eventVthing \text{ } X])]$
- The fourth sense captures the notion of **exchange** (in the hierarchy ‘choice and exchange’, section 4.2) : litt.: *I substitute my hors d’oeuvre against (=for) a dessert*, representation is as follows:  
 $\lambda X, \lambda Y [path \text{ } EXCH_{+poss}([thingVevent \text{ } X], [thingVevent \text{ } Y])]$ .
- The last sense is related to the expression of the **ratio or proportion** (hierarchy ‘quantity - proportion or ration,): litt. *9 votes against 12*:  
 $\lambda X [amount \text{ } AGAINST_{+quant}([amount \text{ } X])]$ .

As can be seen, representations are all substantially different. Substitutions on basic fields, in particular semantic fields, allow for the taking into account of numerous regular metaphorical uses within a sense.

#### 4.3.2 Preposition and verbs: how to underspecify representations.

The verb-preposition interactions are particularly complex and difficult to characterize. Let us illustrate it here on a simple example with movement verbs.

If we take a verb like *run*, its underspecified representation can be represented as follows, where the object PP is typed to be a path, by means of the typed variable P, without specifying any further detail, since the path gets its representation from the PP:

$$\lambda I, P: [\text{path}], [\text{event } \text{CAUSE}([\text{thing } I], [\text{event } \text{GO}_{+\text{loc}}([I], P)])]$$

If we consider a proposition such as *run towards the river*, where the representation of *towards* is:

$$\lambda I, [\text{path } \text{TOWARDS}([\text{thing v place } I])],$$

the combination of the PP with the verb is possible because the expected type of P:  $[\text{path}]$  subsumes the representation of the preposition *towards*, and, therefore, the representation of the PP. The combination of the verb with the PP is then, with the subject variable left opened:

$$\lambda I, [\text{event } \text{CAUSE}([\text{thing } I], [\text{event } \text{GO}_{+\text{loc}}([I], [\text{path } \text{TOWARDS}([\text{thing v place } \text{river}])]))])]$$

A more subtle situation can be illustrated by the sentence: *push against the wall* which should be ruled out because the type of the PP introduced by *against* is a place, not a path. In fact, it is perfectly comprehensible. The reason is that *push* has a by default path incorporated, which is, roughly speaking, made explicit when it is not realized in the object PP. A more comprehensive representation could be then:

$$\lambda I, P: [\text{path}], [\text{event } \text{CAUSE}([\text{thing } I], [\text{event } \text{GO}_{+\text{loc}}([I], P)])]$$

*By – Default* ( $P, \lambda K, [\text{path } \text{TO}_{+\text{loc}}([\text{place } K])]$ ).

Considering the default option allows an argument of type place, with a by default trajectory implemented by TO (any other primitive denoting a movement could have been used instead). The representation of the sentence is then:

$$\lambda I, [\text{event } \text{CAUSE}([\text{thing } I], [\text{event } \text{GO}_{+\text{loc}}([I], [\text{path } \text{TO}([\text{place } \text{NEXT} - \text{TO}_{+\text{loc}}([\text{place } \text{wall}])])])])]$$

**4.3.3 The Umiacs preposition database.** A very valuable and both practically and theoretically sound database was developed about 8 years ago by Bonnie Dorr, it is accessible at:

<http://www.umiacs.umd.edu/Abonnie/AZ-preps-English.lcs>. This is a very large database of preposition semantic representations, characterized by their LCS representation and, quite often, by a thematic grid. There are about 500 entries, for probably all English prepositions. Each preposition sense in Bonnie Dorr's work receives a comprehensive semantic representation in LCS. Senses are paraphrased by an example, in the spirit of

the WordNet synsets. Some restrictions are added, and syntactic positions are made explicit.

## 5. Prepositions and multilingualism

A major problem with prepositions is their status over different languages. As advocated in the introduction of this chapter, there are languages that do not have prepositions or make little use of them. For example, they use case marks instead. Prepositions may also be used for different purposes, e.g. in verb particle constructions in English, for the partitive construction with *de* in French, or they may also enter into clitic constructions, as e.g. in Hungarian and in Turkish.

In languages that use prepositions, regularities over languages are relatively minor, even for closely-related languages in the same family, and even in concrete and well-mastered domains such as time or space. When looking at a bilingual dictionary it is easy to note that preposition translations are very complex, often involving semantic considerations, not to cite the large idiosyncratic variations. This means that, for example, a EuroWordNet for preposition uses is not for the near future.

Let us give a relatively simple illustration involving French, English, German and Spanish. The abstract sense of ‘VIA’ (via something) is lexicalized in different ways according to relatively generic constraints. It is, in general, realized as *par* in French, and as *durch* in German if there is movement and by *aus* if there is none (*aus dem Fenster*). If the VIA indicates also a kind of means, then *auf* is used (*er kam auf dem schnellsten Weg hierher*). In English, this distinction is approximately realized by respectively *through* and *by way of* (e.g. *he went out through / by way of the window*).

An interesting case is where the preposition is realized (or incorporated) in the verb semantics. For example we use e.g. the preposition *via* in English to say: *she comes from Victor Hugo via her mother*, in French we use *par*: *elle descend de Victor Hugo par sa mère*, and in German there is a complex combination with *ab+stammen* (=abstammen): *Mütterlicherseits stammt Sie von Victor Hugo ab*. (From the mother side she originates from Victor Hugo).

Finally, similarly to French, Spanish generally uses a single preposition *por*: *por la ventana* (through the window), but this preposition has additional uses not accepted in French: *caminar por las calles* (walk in the streets), where *por* is translated as e.g. *dans* (in), with no idea of changing location.

## 6. Overview of the book structure

This book is organized in a very classical way, starting with considerations about lexical aspects of prepositions, then descriptive aspects of their syntax, followed by formal aspects. The book ends by semantic features of preposi-

tions, empirical as well as formal. The themes of this book lie, not surprisingly, at the intersection of linguistics, cognitive science, computational linguistics and artificial intelligence.

The book starts with an introduction (this chapter) to the syntax and semantics of prepositions, where basic knowledge about this category is presented. This chapter is essentially designed for novices in the field.

The first group of papers is devoted to the lexical aspects of prepositions: what is their status, how they are represented and how multilingual aspects can be treated. The first paper, by *Luc Baronian*, introduces a lexical portmanteau description of Quebec French prepositions, which is particularly developed compared to standard French. Accounting for the distribution of such prepositions is quite tricky, and a blocking by category analysis is proposed, that can be incorporated in many frameworks (e.g. the Abeillé et al. in this volume). The next paper, by *Andrew McMichael* investigates how English prepositions and adverbial particles are genetically related from a morpho-syntactical perspective. For example, it is shown that many spatial adverbs are reduced prepositional phrases, thus confirming the view of some grammars that equate the PP with an adverb. The next paper, by *David Stringer*, focuses on the role of PPs in directional predication across languages. Starting from Talmy's typology of verb-framed and satellite framed languages, and based on French, Japanese and English, the author shows that these three languages do conform to universal syntactic principles, despite typological differences. It is argued that prepositions form a closed class, the elements of which combine with each other and with spatial nouns to form complex PPs. This first part ends by a contribution by *Mikel Lersundi et al.*. This article describes a common inventory of interpretations for postpositions in Basque and prepositions in English and Spanish. The inventory is a flat list of tags, based mainly on thematic roles. Using the same inventory allows for the identification, for each postposition or preposition, of its translation depending on its interpretation.

The second group of papers focuses on foundational and general aspects of the syntax of prepositions. A first paper, by *Martin Volk*, focuses on the well-known problem of PP-attachment. Attachment tendencies in German of contracted prepositions, pronominal adverbs and reciprocal pronouns are investigated. A statistical method is developed to resolve PP attachment based on unsupervised learning and then an evaluation procedure against a gold standard test set is proposed. The next paper, by *Markus Kracht*, focuses on directional and non-directional locatives and the question of whether they are to be interpreted directionally or not. This paper establishes that locative cases have a layered structure. It also shows that head selection is not case selection but directionality selection. The last paper, by *Aline Villavicencio*, is of a different nature. It investigates ways of using the web to help validate and filter candidate phrasal verbs, in particular those automatically extracted from

corpus of verbs that productively form combinations with particles. Techniques presented in Baldwin (this volume) can be applied to these phrasal verbs to determine whether they are compositional or not. The technique proposed by the author can also be used in M. Volk's contribution (this volume).

The third part of the book is devoted to the insertion of syntactic aspects of prepositions within formal frameworks, in particular HPSGs. Some connections with semantics are also advocated and implemented, such as the reference to MRS (Minimal Recursion Semantics). The first paper, by *Valia Kordoni*, introduces a robust and deep analysis of indirect prepositional arguments in a multilingual context (Greek, English, German) using HPSGs and MRS. The treatment proposed can be coupled with language generation analysis of PPs. This paper has direct relations with the other HPSG papers: Abeillé et al., and Baldwin et al (this volume) and also with Jensen et al., and Mari et al. chapters devoted to ontological characterizations of restrictions holding over prepositions. The next paper, by *Anne Abeillé et al.*, presents, within the HPSG framework, a very comprehensive descriptive overview of the uses of the very polysemic French prepositions *à* and *de* and the properties of the constructions they appear in. The complexity of the data argues against a unitary syntactic and/or semantic treatment, but the empirical facts are nevertheless organized in a systematic fashion. The next paper, by *Tim Baldwin et al.*, the authors outline some of the syntactic and semantic idiosyncrasies of determinerless PPs in English and Dutch, proposing several HPSG-based analysis that capture different subclasses of this phenomenon. The fourth paper, by *Beata Trawinski et al.*, deals with compositional aspects of a variety of PPs consisting of a preposition a noun, another preposition and an NP. Lexicalized Flexible Ty2, which is an adaptation to HPSG of flexible Montague Grammar is considered for that purpose. The last paper, by *Tim Baldwin*, refutes the conventional wisdom that vector-based models are not suited for modelling preposition syntax and semantics. The author also provides empirical evidence for a divergence in the sense of transitive and intransitive usages of prepositions.

The fourth part of the book is devoted to semantics aspects of prepositions. This part covers a large variety of aspects, which are not proper to this category, among which: semantic classifications, ontologies, polysemy, and semantic formalisms, including Lexical Conceptual Structure and Discourse Representation Theory. Some of these papers are theoretically oriented while others show how research on prepositions can benefit to Computational Linguistics. The first paper of this group, by *J. Kelleher et al.*, proposes a computational modelling of the spatial semantics of projective prepositions, with particular attention to frames of reference ambiguity and object occlusion. Within our cognitive grammar, the authors have developed a novel algorithm for locating the spatial templates origin, a scalable 3D model of projective spatial templates that integrates the impact of the perceptual cue of object occlusion and finally,

building on previous psycholinguistic work, they have developed an algorithm for handling the impact of frame of reference ambiguity on the construction of spatial templates. The next paper, by *Per A. Jensen et al.*, investigates a formal ontological semantics for modelling the conceptual contents of NPs with focus on the contribution of embedded PPs. The framework posits relationships between prepositions and semantic roles, thereby accounting for their polysemous semantics. Ontological wellformedness conditions are imposed on the ontology in the form of so-called ‘affinities’ enabling disambiguation of NPs containing PPs and similar constructions such as Noun-Noun compounds and genitive constructions. The next paper, by *Ryusuke Kikuchi et al.*, proposes a context dependent interpretation of the Japanese postposition ‘No’ within the SDRT (Segmented Discourse Representation Theory). The meaning of this postposition corresponds roughly to the preposition “of” and the possessive marker “s” in English. Typically, *no* forms a noun phrase,  $NP_1 \ no \ NP_2$ . Interestingly, the meaning of this noun phrase depends not only on the semantic properties of NPs occurring in the construction but also on contextual information. The forth paper, by *Alda Mari*, presents an analysis and a formal model for the notions of instrumentality and manner through the study of the preposition *avec*. These two notions are very often assimilated to one another. The author investigates the semantic foundations of this intuition, and proceeds via a bottom-up analysis, considering first the meanings-in-context and then the underspecified formal representation common to both of these notions. The Fifth paper, by *Alda Mari et al.*, presents an concrete analysis and a formal model based on the Lexical Conceptual Structure for Prepositions denoting instrumentality, based on corpus studies in French. This paper concludes by a generic, underspecified model for the abstract notion of instrumentality. The last paper, by *Farah Benamara et al.*, shows how prepositions are used in WE-BCOOP, a logic based cooperative question answering system. The authors focus on a subset of French spatial and temporal prepositions. An adequate interpretation in terms of Euclidean geometry is proposed. Then, the authors show how these representations are used in specific reasoning schemas such as conceptual relaxation, based on a set of relations that classify each preposition according to its interpretation. Finally, the authors give some hints on how prepositions can be generated in natural language both during the aggregation and the lexicalisation phases proper to natural language generation.

## References

Baker, M.C., (1988), *Incorporation: A Theory of Grammatical Function Changing*, Chicago University Press.

Boguraev, B., Sparck Jones, K., (1987) A Note on a Study of Cases, Computational Linguistics, vol. 13, 1-2, p. 65.

Cannesson, E., Saint-Dizier, P.,(2001), *A general framework for the representation of prepositions in French*, ACL01 WSD workshop, Philadelphia.

Cervioni, J., (1991), *La préposition: Etude sémantique et pragmatique*, Duculot, Paris.

Cruse, A., (1973), Some Thoughts on Agentivity, *Journal of Linguistics*, vol 9-1.

Cruse, A., (1986), *Lexical Semantics*, Cambridge university Press.

Dorr, B., Olsen, M.B., (1997), *Deriving Verbal and Compositional Verbal Aspect for NLP Applications*, proc. ACL'97, Madrid.

Dorr, B., (1993), *Machine Translation, a view from the lexicon*, MIT Press.

Dorr, B. J., Garman, J., and Weinberg, A., (1995), *From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT*, Machine Translation, 9:3-4, pp.71-100.

Dorr, B., Jones, D., (1996), *Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues*, in proceedings of *Coling 96*, Copenhaguen.

Dowty, D., (1989), *On the Semantic Content of the Notion of Thematic Role*, in G. Cherchia, B. Partee, R. Turner (eds.), *Properties, Types and meaning*, Kluwer Academic.

Dowty, D., (1991), *Thematic Proto-roles and Argument Selection*, Language, vol. 67-3.

Fellbaum, C., (1993), *English Verbs as Semantic Net*, *Journal of Lexicography*, vol. 6, Oxford University Press.

Fillmore, C., (1968), *The Case for Case*, in *Universals in Linguistic Theory*, E. Bach and R.T. Hams (eds.), Holt, Rinehart and Winston, New York. Gruber, J., (1976), *Studies in Lexical Relations*, MIT doctoral dissertation and in *Lexical Structures in Syntax and Semantics*, North Holland (1976).

Horno Chéliz, M.,del C., (2002), *Lo que la preposición esconde*, University of Zaragoza press.

Jackendoff, R., (1972), *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge.

Jackendoff, R., (1983), *Semantics and Cognition*, MIT Press, Cambridge.

Jackendoff, R., (1987), *The Status of Thematic Relations in Linguistic Theory*, *Linguistic Inquiry*, vol. 18.

Jackendoff, R., (1990), *Semantic Structures*, MIT Press.

Levin, B., (1993), *Verb Semantic Classes: a Preliminary Investigation*, Chicago University Press.

Lindstromberg, S.,(1997), *English Prepositions Explained*, John Benjamins.

Mari, A.,(2003), *Polysémie et Décidabilité. Le cas de avec ou l'association par les canaux*, Paris: L'Harmattan.

Moriceau, V., Saint-Dizier, P.,(2003), *A Conceptual Treatment of Metaphors for NLP*, in proc. ICON'03, Mysore.

Pesetsky, D., (1982), *Paths and Categories*, MIT doctoral dissertation.

Pinker, S., (1989), *Learnability and Cognition: The acquisition of argument structure*, MIT Press.

Pugeault, F., Saint-Dizier, P., Monteil, M.G., (1994), *Knowledge Extraction from Texts: a method for extracting predicate-argument structures from texts*, in proc. Coling 94, Kyoto.

Pinkal, M., (1985), *Logic and Lexicon*, Oxford : Oxford University Press.

Pustejovsky, J., (1991), *The Generative Lexicon*, Computational Linguistics, vol. 17, MIT Press.

Pustejovsky, J., (1995), *The Generative Lexicon*, MIT Press.

Rappaport, M., Levin, B., (1988), *What to do with  $\theta$ -roles?*, in *Syntax and Semantics 21: Thematic Relations*, W. Wilkins (ed.), Academic Press.

Ravin, Y., (1990), *Lexical Semantics without Thematic Roles*, Oxford Univ. Press.

Reuland, E., Abraham, W., (eds.), (1993), *Knowledge and Language, Vol II*, Kluwer Academic.

Roca, I.M. (ed.), (1992), *Thematic Structure: its Role in Grammar*, Mouton de Gruyter, Berlin.

Saint-Dizier, P. (1999), *Alternations and Verb Semantic Classes for French*, in Predicative Forms for NL and LKB, P. Saint-Dizier (ed.), Kluwer Academic.

Saint-Dizier, P., (2005), PrepNet: a Framework for Describing Prepositions, preliminary investigation results, IWCS, Tilburg.

Talmy, L. (1976), *Semantic Causative Types*, In M. Shibatani (ed.), *Syntax and Semantics 6: The Grammar of Causative Constructions*. New York: Academic Press, pp. 43-116.

Talmy, L., (1985), *Lexicalization Patterns: Semantic Structure in Lexical Forms*, in *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon*, T. Shopen (ed.), 57-149, Cambridge University Press.

Wierzbicka, A. (1992), *Semantic Primitives and Semantic Fields*, in A. Lehrer and E.F. Kittay (eds.), *Frames, Fields and Contrasts*. Hillsdale: Lawrence Erlbaum Associates, pp. 208-227.

Wilks, Y., (1977), *Good and Bad Arguments about Semantic Primitives*, Communication and Cognition, vol. 10, pp. 181-221.

Woods, F.T., (1979) *English Prepositional Idioms*, Macmillan, London, UK.

## Chapter 2

# PREPOSITION CONTRACTIONS IN QUEBEC FRENCH

Luc V. Baronian

*Stanford University  
Linguistics Department  
Stanford, CA 94305-2150 USA  
baronian@stanford.edu*

**Abstract** Contractions of prepositions and articles (P+A) in Quebec French (QF) are analyzed in this paper as portmanteau prepositions with definiteness, number and gender features. Examples of P+A are [dē] for *dans les* and [a:] for *à la*, and have the same status, I argue, as the older *au* and *du* of Standard French. The core part of the paper gives arguments for this analysis from sociolinguistics, speech style, orthography, phonology and syntax. The implications of this analysis for the grammar of QF and eventual computational accounts that would want to generate the uses of the portmanteau vs. the simple prepositions are then examined. One implication is phonological: we have to admit a new phoneme in the language, /a:/. This is an interesting result, as the appearance of this new phoneme and new portmanteau prepositions seems to correspond in time to the penetration into the system of three other long vowels from English loanwords. Secondly, in order to explain the distribution of these new portmanteau prepositions, an account is proposed based on category blocking of the regular prepositions by the selectional properties of the portmanteau prepositions. This account has the advantage of also explaining a previously unnoticed gap in the combination of prepositions and articles in QF.

**Keywords:** Portmanteau prepositions, Quebec French, syntax, coordination, phonology, vowel fusion, liquid deletion, blocking, syntactic gap

## 1. Introduction

This paper looks at a phenomenon in Quebec French (QF) that I term preposition contractions or simply contractions, for short. It is the surface contraction of a preposition and an article (P+A) as one element from which the original two are not readily distinguishable anymore. After introducing in §2 the

contractions specific to QF and some other contractions shared with Standard French (SF), I argue in §3-5 that most of them cannot be treated in the phonology, as has been previously assumed by authors such as (Dumas, 1974) and (Walker, 1984). The evidence presented there will lead me to conclude that most QF P+A contractions have been lexicalized, just like the contractions shared with SF. (Santerre et ali., 1977) had suggested that the QF P+A contractions are probably lexicalized without however weighing the arguments on both sides and more recently (Tremblay, 1999) implies that the QF ones need the same treatment as the ones shared with SF.

In the conclusion (§6), I examine two implications of the analysis of QF P+A as lexicalized function words for the grammar of QF. The first implication is phonological: we have to admit the existence of a new underlying vowel, /a:/, which occurs only in the new portmanteau prepositions. It seems that the new portmanteau prepositions and this new phoneme have entered the grammar in the same period as three other long vowels from English borrowings. Secondly, since the set of such portmanteau prepositions has been extended, the new pattern will require some explanation. They will be explained through category blocking in the syntactic selection properties of prepositions, a solution that will have the advantage of also explaining a gap that has remained unnoticed until now, as far as I know. Consequences for theories that try to generate just the right uses of contracted vs. simple prepositions will also be discussed.

In this paper, contraction is used as a descriptive term. It is also a neutral term in two respects: 1) neutral as to whether the “contraction” is diachronic (has happened at a point in time and speakers do not contract them on-line) or synchronic (speakers contract them on-line); 2) neutral as to whether the contraction is or was phonological or morphological.

Five native speakers of QF other than myself were consulted for this study in December 2000. Two (a female, 57; a male, 51) were born in Abitibi. The female lived most of her life in Montreal, although she also lived for about ten years as a child in the Ottawa Valley. The male lived a large part of his life in Quebec city. The three other speakers lived most of their lives in the Ottawa Valley (two females, one 45, one 25; one male, 44). My own judgements (male, 26, Montreal) were also considered.

## 2. Facts

### 2.1 French P+A

In French, the prepositions *à* and *de* can be considered as functional prepositions (as opposed to lexical prepositions): they have more general semantic content, they are less free syntactically, they are unstressed and they combine morphophonologically with some definite articles. (Tremblay, 1999)

argues that it is not accurate to claim that they are devoid of any semantic content, because in some uses they clearly are associated with a meaning:

(1) a. Je vais à l'école.  
I go to school. b. La pomme à Jean.  
The apple (belonging) to John.

In (1a), the associated meaning is “destination/location” and in (1b) it is “possession”. However, they do have semantically opaque uses such as (2), where they serve more argument marking purposes, and uses such as (3b), and to a lesser extent (3a) (a ‘source/location’), overlapping with (1), for certain speakers.

(2) a. Je le promets à Michel.      b. La construction de la maison.  
I promise it to Michel.              The construction of the house.

(3) a. Je reviens de l'école.              b. Le livre de Michel.  
I am coming back from school.      The book (belonging) to Michel.

One example of the “less free” syntactic status of these two prepositions in QF is the fact that their complement cannot be omitted at the end of simple sentences, contrary to the lexical prepositions:

(4) a. Je sors avec (elle). b. Je reviens de <sup>\*</sup>(l'école).  
I am going out with (her). I am coming back from <sup>\*</sup>(school).

In most varieties of French I am familiar with, neither *à* nor *de* can occur followed by the masculine singular definite article *le* or the plural definite article *les*. Instead, one uses respectively the morphonologically fused contractions *au/aux* for *à* and *du/des* for *de*. These are summed up in (5). Note that if the following word is singular and starts with a vowel, there is no contraction, even if the word is masculine, and note also the z-liaison with the plural contractions when the following word starts with a vowel.

## (5) Standard French Preposition + Definite Article sequences

a.	Pre V sing.	à l'	[al]	b.	Pre V sing.	de l'	[dəl]
	Fem. sing.	à la	[ala]		Fem. sing.	de la	[dəla]
	Masc. sing.	au	[o]		Masc. sing.	du	[dy]
	Plural	aux	[o(z)]		Plural	des	[de(z)]

Not using the contracted forms in the required context yields strong ungrammaticality judgements for any speaker of both SF and QF (cf. 6). The only apparent counterexample to this claim is in the context where the article is part of a fixed expression. In these cases, there is cross-dialectal and inter-speaker variation. For example, we can see in (7) that most QF speakers prefer to use the contractions in lexicalized NP cases, although both realizations are possible.

(6) La tarte {aux}/\*{à les} bleuets. ‘Blueberry pie’

a. Ce film-là ressemble {aux}/\*{à les} Misérables.  
This movie ressembles Les Misérables.

(7)

b. Je déteste le début {des}/\*{de les} Misérables.  
I hate the beginning of Les Misérables.

The use of these contracted forms is many centuries old and it should be rather uncontroversial that they aren't contracted on-line by speakers. Indeed, it is more economical to admit them as lexicalized prepositions with definiteness, number and gender features, rather than to complicate the phonology for just a handful of forms. If this analysis is right, the unidirectional grammaticalization hypothesis of (Hopper & Traugott, 1993) is appealing to explain the fact that they haven't been “decontracted” in any dialect. It is especially striking that they have been stable for so long since they have been reported to be acquired rather late by children,<sup>1</sup> so one would expect some non-standard dialects to have undone the contractions, but this is apparently not the case, a fact supporting a theory of unidirectional grammaticalization.<sup>2</sup>

## 2.2 QF P+A

QF shows on the surface additional P+A contractions. In (8), the reader can compare the QF surface forms in the same contexts as in (5) for SF. Note that all contractions present in SF are still present in QF. The contractions not shared with SF are in bold characters.<sup>3</sup> (The prepositions *à* and *de* do not show clear contractions on the surface with indefinite articles).

(8) Contractions of *à* and *de* with definite articles

a. Pre V sing.	à l'	[al]	b. Pre V sing.	de l'	[dəl]
Fem. sing.	à'	[a:]	Fem. sing.	dela~ <b>da</b>	[dəla]~[da]
Masc. sing.	au	[o]	Masc. sing.	du	[dy]
Plural	aux	[o(z)]	Plural	des	[de(z)]

More interesting is the fact that such contractions have arisen with two other prepositions,<sup>4</sup> *dans* ‘in’ and *sur* ‘on’, as shown in (9).<sup>5</sup> Because these two prepositions showed no contraction in previous stages of French, there are more new contractions with them than with *à* and *de*, since the latter were already contracted in the plural and masculine singular. Even more interesting is the fact that, unlike the cases of *à* and *de*, clear contractions have arisen between these prepositions and indefinite articles (10).<sup>6</sup> Note that (10a) are contractions of *dans*, not *de*.

(9) Contractions of *dans* and *sur* with definite articles

a.	Pre V sing.	dans l'	[däl]	b.	Pre V sing.	su(r) l'	[sy(ʁ)l]
	Fem. sing.	<b>dans'</b>	[dā]		Fem. sing.	<b>s'a</b>	[sa:] or [sqa]
	Masc. sing.	dans le	[däl]		Masc. sing.	su(r) le	[sy(ʁ)lə]
	Plural	<b>dins</b>	[dē(z)]		Plural	<b>s'es</b>	[se:(z)] or [sue(z)]

(10) Contractions of *dans* and *sur* with indefinite articles

a.	Pre V sing.	<b>d'un</b>	[dōn]	b.	Pre V sing.	<b>s'un</b>	[sōn]
	Fem. sing.	<b>d'une</b>	[dyn]		Fem. sing.	<b>s'une</b>	[syn]
	Masc. sing.	<b>d'un</b>	[dō]		Masc. sing.	<b>s'un</b>	[sō]
	Plural	dans des	[dāde]		Plural	su(r) des	[sy(ʁ)de]

## 3. Two possible analyses

## 3.1 Analysis 1: Phonology

(Dumas, 1974) and (Walker, 1984) do not discuss the facts presented in §2.2 directly, but group them under analyses of the phonology of QF. In (11-12), I give a simplified version of their phonological account. On the left, I give a (very) simplified version of the two sequential phonological rules they consider to be at work here: a (lexical) rule of inter-vocalic liquid deletion that applies only to some clitics (pronouns and articles) and a (post-lexical) rule of vocalic fusion. On the right, I illustrate the derivation with the underlying sequence *dans les*.

dans les [dāle]

(11) Liquid Deletion (LD)  
l, r → Ø/V<sub>1</sub>-V<sub>2</sub> (In some clitics). [dāe]

(12) Vowel Fusion (VF)  
V<sub>1</sub>V<sub>2</sub> → V<sub>2</sub>: (Output nasal if V<sub>1</sub> or V<sub>2</sub> is nasal). [dē:]

So for example, Rule (11) is meant to also account for the inter-vocalic liquid deletion (LD) in (13a), which is blocked in (13b) because of the preceding consonant:

(13) a. V-V: Je crois la fille. b. C-V: J'aime la fille.  
[ʃkruwa a fil] [ʒem la fij]  
I believe the girl. I like the girl.

As for Rule (12), its effect can be observed with any two adjacent vowels. It is labeled in the literature as vowel fusion (VF) and it is an automatic post-lexical phenomenon. (14) gives two examples of the effect of VF in contexts other than clitics.

(14) a. chocolat au lait 'chocolate milk' b. vraiment heureux 'really happy'  
/ʃkɔla o la/ [ʃkɔl o: la] /vraemā ɔʁø/ [vraem ɔ: ʁø]

### 3.2 Analysis 2: New portmanteau prepositions

Analysis 2, which I am proposing, hypothesizes that most QF P+A contractions in bold in (8) through (10), are new portmanteaux elements, like the older *au*, *aux*, *du* and *des*. What I propose, more precisely, is that LD and VF have applied sequentially in most P+A at some point in the history of QF and that the P+A have since been reanalyzed as a single unit.

Without further evidence, by Occam's razor, since the mechanisms used to account for LD and VF seem to be independently needed synchronically (whether as rules, constraints, etc.), Analysis 1, or a modern constraint-based version of it, should be considered as essentially on the right track and Analysis 2 should be abandoned. In §4-5, however, I will provide evidence that shows how the latter analysis wins over the former, except for some speakers in the case of the indefinite singular articles only.

## 4. External evidence for Analysis 2

### 4.1 Sociolinguistic evidence

In a sociolinguistic study on LD in QF, (Santerre et al., 1977:536) noted that upper class speakers tend to avoid LD, but that they nevertheless delete many liquids, especially in sequences that the authors say must be lexicalized, such as [sa:], [dɛ], etc., i.e. P+A contractions.

Whether or not we accept Santerre et al.'s suggestion that P+A contractions are lexicalized, one thing is clear from their study: LD, if it is happening in P+A, must have a different status in P+A than in other instances such as (13a), at least for upper class speakers, since they avoid it less in P+A sequences than elsewhere.

### 4.2 Staccato speech

A stronger argument in favor of Analysis 2 would be if we could find some styles where LD and VF are not observed, but the P+A sequences still show up contracted. Speakers, of course, cannot reliably tell in which styles they produce this or that linguistic phenomenon.

In order to do so then, I investigated a style, staccato speech, which has the advantage of providing contexts where speakers can make grammaticality judgements about the possibility of using LD and VF or not, rather than simply giving their opinion on whether they would prefer to use them or not in this style.<sup>7</sup>

Staccato speech is a style where one inserts small breaks or pauses between syllables. This style is used to “weigh one's words”, that is, to insist on a part of a sentence, for example, when speaking to children. In (15), we see that some P+A contractions occur in staccato speech (15a), but true cases of VF

(15b) and LD (15c), cannot be realized in this style. (The dots ● mark the pauses).

(15) a. P+A (definite article)  
Check ● dins ● boîtes  
Check in the boxes.

b. VF  
\*[ʃɔ ● kɔ ● lo: ● le] chocolat au lait  
[ʃɔ ● kɔ ● la ● ɔ ● le] ‘chocolate milk’

c. LD  
Je ● crois ● [la]/\*[a] ● fille.  
I believe the girl.

I conclude from this that speakers are not using whatever mechanism it is they are using in VF and LD cases in their P+A contractions. This result supports Analysis 2. However, the P+A contractions with singular indefinite articles don’t pass the staccato test (16) for most speakers.

(16) P+A (indefinite article)  
La ● pou ● pée ● ?\*{[dyn]}/{dans ● une} ● boîte.  
The doll in a box.

### 4.3 Orthography

Although there is no fixed orthography for colloquial QF, there is an abundant literature that can be consulted where either dialogues, popular songs or old folktales are transcribed using an improvised orthography to render the phonological differences from SF. Actually, this orthography is not always improvised, it is semi-conventionalized, as the same spellings often appear in different authors’ works. Many of the P+A contractions have found regular spellings (with more or less variation) in the literature and in email exchanges: *dins* for *dans les*; *à’* for *à la*; *dans’* or simply *dans* for *dans la*; *s’a* for *sur la*; etc.

On the other hand, I know of no other cases of VF that have developed such conventional spellings, although LD is also often noted.<sup>8</sup> In order to provide a more serious argument than these personal observations, I systematically looked for such popular spellings in the song booklet of the CD (Desjardins, 1993). There, I found that every P+A contraction is transcribed as pronounced, except one,<sup>9</sup> but that no other cases of VF are transcribed as such. This finding is in line again with Analysis 2 and goes against Analysis 1.

## 5. Core linguistic arguments for Analysis 2

### 5.1 Syntax: NP and PP Coordination

In this section, I show how the QF P+A contractions behave like the older contractions shared with SF. This behavior will turn out to be inconsistent with the phonological analysis, but is of course to be expected under Analysis 2.

First, observe the sentence in (17), where two PPs are coordinated. Notice how in (18) the preposition can select for two coordinated NPs.

(17) Regarde [dans le placard] pis [dans la cave].  
 Look [in the wardrobe] and [in the basement].

(18) Regarde dans [le placard] pis [la cave].  
 Look in [the wardrobe] and [the basement].

Notice now that if a sentence as in (19) uses a preposition of SF with definiteness, number and gender features (I will henceforth call these portmanteaux, for portmanteau prepositions), extraction of the prepositional element is impossible, as shown in (20).

(19) Regarde [aux portes] pis [aux fenêtres].  
 Look [from the doors] and [from the windows].

(20) \*Regarde aux/à les portes pis les fenêtres.  
 Look from [the doors and the windows].

This behavior is paralleled by the new portmanteaux of QF in (21) and (22):

(21) Regarde [dins placards] pis [dans cave].  
 Look [in the wardrobes] and [in the basement].

(22) ?Regarde dins placards pis la cave.  
 Look in the cupboards and the basement.

Notice that under the phonological analysis, since it is possible to extract a preposition (17-18), it should be possible to extract the preposition in (21), LD and VF then applying to give us (22). Some speakers do accept (22) at first, but upon further investigation, they mention they've heard the construction, but wouldn't use it. Further, every speaker I consulted without exception clearly told me that (21) is the way to express this sentence. As we will see in §6, the grammaticalization of the P+A sequences seems to be fairly recent (XXth century). In this situation, it is to be expected that some speakers might have heard, as they say, such sentences as (22) from older speakers who have not reanalyzed the P+A sequences. It is also possible that they interpret (22) as a syntactically standard sentence with colloquial phonology. In either case, it is expected that (22) would be less grammatical.

Another similarity is that plural portmanteaux can be extracted (23) from a coordination of plural nouns for some speakers if the coordinated nouns are a fixed expression or are closely related semantically:

(23) Priorité aux [femmes et enfants].  
 Priority to [women and children].

Extraction of the whole portmanteau from a coordination including at least one singular noun yields ungrammaticality:<sup>10</sup>

(24) \*Je reviens du [dépanneur et bureau de poste].  
I'm back from the [corner store and post office].

The new QF portmanteaux again parallel the behavior for some speakers (25). Some speakers don't like (25a), but that could just mean that *café et bistro*s isn't a common expression for them, the crucial point is that (25b) was clearly ruled out for all speakers, while (25a) was acceptable at least for some speakers.

(25) a. ? Je sors dins [cafés et bistro].  
I go out in the [cafés and bistro].  
b. \*Je mange dans' [cuisine et salle à manger].  
I eat in the [kitchen and dining room].

We thus have evidence that the syntactic behavior of new P+A contractions is similar if not identical to that of the portmanteau elements shared with SF. Finally, if you recall (7b-c), QF speakers usually prefer to use SF P+A contractions even when the article is part of a lexicalized NP. In (26a-b), we can see that they also prefer to use the P+A specific to QF rather than the non fused variants. The SF equivalents without LD or VF are also of course accepted by all speakers.

(26) a. Il jouait {?[dã ez]}/{[dẽz]}/{[dã lez]} Oraliens.  
He played in Les Oraliens (television show).  
b. Mets donc ça {?[dã e]}/{[dẽ]}/{[dã le]} boîtes.  
Why don't you put that in the boxes.

## 5.2 Phonetics/Phonology

Ironically, the final nail in the coffin of Analysis 1, which was a phonological one, comes from phonological or phonetic evidence. In QF, underlying long vowels undergo shortening in pre-stressless position, cf. (Meillet, 1912/1958:138) for SF and (Paradis & Deshaies, 1991) and (Ouellet & Thibault, 1996) for QF:

(27) V : → V/\_  $\bar{\sigma}$

The rule in (27) should be taken as a descriptive statement, good enough for our purposes. Perhaps there is rather a shortening in stressless positions (from a constraint forbidding long stressless vowels) and that when adjacent to a stressed syllable, long vowels tend to attract stress, which allows them

to keep their length. In such positions, P+A contractions undergo shortening, while long vowels arising from true VF do not undergo it, as (28c) shows us.<sup>11</sup>

(28) a. Pre-stressless P+A [dĕ] fi'lets in the nets b. Pre-stress P+A [dĕ:] 'bars in the bars

c. Pre-stressless VF Less rues à Montréal. { [ra:] } / { \* [ra] } d. Pre-stress VF rues à Rouyn. [ra:] 'The streets in Montreal.' 'The streets in Rouyn.'

The conclusion to draw from (28) is that there is a boundary in the middle of true VF cases, which isn't there in P+A constructions. This is straightforwardly accounted for under Analysis 2, but again remains unexplainable under Analysis 1.

## 6. Conclusion: consequences of the analysis

The reanalysis of clitic contractions by VF as new portmanteaux has two implications for the grammar of QF, each raising some interesting theoretical questions. The first implication is phonological: a new vowel (the 21st by my count), /a:/, must be admitted. This new vowel has already given rise to minimal pairs:

(29) à /a/ 'at'  
 sa /sa/ 'POSS-FEM-SG' à /a:/ 'at-FEM-SG-DEF'  
 s'a /sa:/ 'on-FEM-SG-DEF'

Further, the reanalysis seems to have taken place sometime in the XXth century, at about the time when words with long vowels not in the system started penetrating the language through English borrowings (e.g., *cheap* /i:/, *cool* /u:/, *steak* /e:/). Note that as a daughter-language of XVIIth century SF, QF already had long vowels in its system: long /ɔ: o: ə:/ opposed to short /ɛ ɔ: œ a/, plus the four nasal vowels, which are long. However, before the XXth century, the English vowels in *bean*, *balloon* and perhaps *steak* were rendered short in borrowings (e.g., *bean* [bɪn], *balloon* [balʊn]).<sup>12 13</sup> These three phonemes are now rendered long [i: u: e:] (or slightly diphthongized) in new borrowings (e.g., *cheap*, *cool*), and we have evidence that they have even penetrated the native stock of words. For example, there was a phonetic lengthening when the next coda consonant was one of the following: /r v z ʒ/ (*douze* ‘twelve’ showed up as [du:z]), but now this length has been carried over to related words (so we get *douzième* ‘twelfth’ [du:zjɛm], as opposed to *cousine* ‘female cousin’ [kuzin]).<sup>14</sup>

Notice how we now have a length opposition for most oral cardinal vowels in the system:

	i /i:	u:/u
(30)	e/e:	o:/...
	ɛ/ɜ:	ɔ:/ɔ
	a/a:	a:/...

However, we now face a chicken and egg problem, as it is not clear which of the three phenomena allowed for the others. (Sound change in words like *douze*, new borrowing strategy from English or reanalysis of P+A sequences yielding a new phoneme). Notice also, that it is not every vowel of English that was suddenly available as a phoneme for QF speakers (e.g., [ʌ] is still rendered as [ɔ]). Unfortunately, investigating this problem would take us too far afield for this paper.

The second implication, from my analysis that we have to admit new portmanteau prepositions with definiteness, number and gender features, is that any syntactic framework aspiring to account for QF must be flexible enough to allow both portmanteaux and non-contracted P+A sequences (as it should already do for SF). However, the same theory must also be able to restrict the occurrence of non-contracted P+A sequences to just the right cases. This second task is a non-trivial one, because if you recall the data in §2.2, the context in question is not as predictable in QF as in SF. Whereas in SF, it is only the plural and preconsonantal masculine definite articles that contract with the prepositions, in QF it is hardly predictable in those terms: sometimes the masculine articles contract, sometimes not, sometimes the feminine articles contract, sometimes not.

One way around this problem is to propose that all combinations of a functional preposition with an article are prepositions with definiteness, gender and number features, even if they happen to look exactly like the syntactic combination of the two parts. For example *dans le* and *sous les*, which don't show phonological contraction like *dans les* /dɛ:/ or *à la* /a:/, would still both be one word, respectively /dāl/ and /swe/. In this way, since whole classes of preposition and article sequences are considered as one word, the non portmanteau prepositions can select for the remaining classes of NPs.

If you recall the facts, it is always with the singular indefinite articles that the functional prepositions prove to be part of a different word. Concretely, my proposal is that the functional prepositions *à*, *de*, *dans* and *sur* (and perhaps also *sous*, *'ec* and *sans*, which, syntactically behave like *à* and *de* in not allowing any argument deletion) would select only for indefinite singular NPs and not for plural or definite NPs. These latter would be selected by the right portmanteau prepositions. We also have to allow the non portmanteau prepositions to select singular proper nouns (*à Pierre* 'to Pierre'), quantified phrases (*à tous les humains* 'to all the humans', *à deux personnes* 'to two people'), possessives (*à mes amis* 'to my friends') and demonstratives (*à ces personnes* 'to these people').

The careful reader will remember that the evidence in (18), where the preposition *dans* can stand outside a coordinated NP with two articles inside, speaks against such a generalization. However, every speaker I consulted prefers to repeat the preposition with the two members of the coordination as in (17). Therefore, it is not an unreasonable hypothesis to assume that the dispreferred realization with a single preposition is a SF one, which every speaker of QF I consulted masters relatively well. In fact, (Larousse, 1964:572) mentions that one usually does not repeat prepositions other than *de*, *à* and *en* in coordinated structures.

An immediate advantage of this proposal is that it explains a gap in the possible combinations of the preposition *de* and articles, which, as far as I know, has never been noticed before. The gap is that it is impossible in QF to express the sequence *de des* (*de*-preposition + indefinite plural article), either as a contraction (31) or as a spelled-out sequence (32). Cf. indefinite singular (33), definite singular (34) and definite plural (35).

- (31) \*La vie des insectes est courte.  
The life of insects is short (indefinite interpretation).
- (32) \*La vie de des insectes est courte.
- (33) La vie d'un insecte est courte.
- (34) La vie de l'insecte est courte.
- (35) La vie des insectes est courte.

Such a case is avoided by speakers simply by paraphrasing the intended meaning otherwise, as in (36):

- (36) La vie de certains/quelques/plusieurs insectes est courte.  
The life of certain/some/many insects is short.

This kind of avoidance is typical of paradigm gaps. For example, the verb *frire* 'to fry' behaves in such a way in the indicative imperfect, where there is such a gap. Speakers are incapable of conjugating *frire* in this tense (37) and tell you they would formulate it periphrastically as in (38).

- (37) La nourriture \*fryait, \*fritait, \*friyait, \*frisait...  
'The food was frying.'
- (38) La nourriture était en train de frire.  
'The food was in the process of frying.'

In SF, it is also impossible to have the sequence *de des*, but instead of having a gap, the language uses simply *de*, which can here be considered a

portmanteau:

(39) La vie d'insectes est courte.

(39), when simply not rejected by QF speakers, strikes them as very formal or literary. One speaker couldn't even get the intended meaning 'The life of some insects is short' and only got the meaning 'Insect life is short' (which, however, would be spelt differently, i.e. with singular *insecte* instead of plural *insectes*).

The other solution I can think of is to allow the portmanteaux to block the free combination of simple prepositions with articles. This has the advantage that we do not have to consider as one word forms such as *dans le*, which does not show decisive evidence for it. However, then, the observed gap in QF could not be maintained, since the default free combination *de des* would occur. A sketch of the two solutions for linguistic theories and computational models is given in (40).

a. **Solution 1: Blocking by category & generalized portmanteaux**

Instead of an adjacent functional preposition and a [PLURAL] or [DEFINITE] article, use the appropriate portmanteau:

[INDEF,PLUR]	[DEF,PLUR]	[DEF,SING] [PreC, M]	[DEF,SING] [PreC, F]	[DEF,SING] [PreV]
/ade/	/o(z)/	/o/	/az/	/al/
	/de(z)/	/dy/	/dəla/	/dəl/
/dāde/	/dē (z)/	/dāl/	/dā/	/dāl/
/syde/	/se(z)/	/syl/	/saz/	/syl/

(40)

In the other contexts, freely combine the relevant prepositions and determiners: *à, de, dans, sur les, le, la, l', des, un, une, ces, ce, cette, cet*, etc.

**Disadvantage:** E.g. *dans le* is one word, though there is no decisive evidence for this.

b. **Solution 2: Blocking word by word**

If one of the following portmanteaux can be used, use it:

/o/, /o(z)/, /az/, /de(z)/, /dy/, /dē (z)/, /dā/, /se(z)/, /saz/

Or else, combine the appropriate preposition and determiner.

**Disadvantage:** The QF gap is erroneously filled by *de des*.

Now note that if the solution in (40a) is the right one, at least the combination of *de* and *à* with all plural and definite articles had to be already ruled out in the stage of QF prior to VF, because it is the only way to maintain a gap. The portmanteau analysis had to be already in place in these cases, even for those that were homophonous with the syntactic combinations. At that stage, the system was almost identical with that of SF, with the exception that SF sim-

ply used *de*, a suppletive form, where QF has a gap, therefore I see no reason to have a different analysis for SF: *à* and *de* cannot combine with plural and definite NPs, instead the portmanteaux *à-des*, aux, *à-la*, *au*, *de*, *des*, *de-la*, *du* must be used, it is just that the form of the SF portmanteaux is more readily predictable than in QF. This explains why a SF grammar such as (Larousse, 1964:572) mentions that these prepositions are repeated in coordinated NPs (some predictable exceptions exist), just like portmanteaux prepositions.

Finally, I take note that the analysis presented in Abeillé et al. (this volume) could shed light on the peculiar distribution of the portmanteaux associated with *de*. If it can be shown that the cases where *de* selects for N' instead of NP independently motivate the same selection in indefinite plural context, then there would be no need to have so many portmanteaux in French.

## Acknowledgments

I thank the five native speakers who agreed to participate in this study, as well as Eve Clark, Paul Kiparsky, Tom Klingler, Will Leben, Yves Charles Morin, Ivan Sag and Arnold Zwicky for helpful comments on previous versions of this paper, and the participants in the TREND 2000 workshop at the University of California, Santa Cruz and the ACL-SIGSEM workshop held by IRIT. I acknowledge doctoral fellowships by Stanford University and the Social Sciences and Humanities Research Council of Canada.

## Notes

1. (Bautier-Castaing, 1977:23) and (Labelle, 1976:62), mention that children still make mistakes at the ages of four and five respectively (I thank Eve Clark for pointing out these references to me).

2. The only French speaking area where I have noticed uncontracted forms, especially with *de*, such as *de le* or *de les* is Louisiana where a French dialect (Cajun) and a French Creole (Louisiana Creole) are spoken. (Neumann, 1985:304-305) notes that the preposition *de* is very rare in the oldest Louisiana Creole texts. I conclude that it is possible that the preposition is a borrowing from French to Louisiana Creole that was generalized. Because of the close interaction between Cajun and Creole in Louisiana, the “decontracted” constructions then got borrowed back into Cajun. (Baronian & Michelet, 2005) show that the constructions without the portmanteaux are significantly more frequent in the speech of Creoles than in that of Cajuns.

3. [da] was reported to me by Yves Charles Morin, although I have never encountered it myself. I don't know at this point if its distribution is social or regional. According to Morin, it is possible that it is restricted to some idiomatic expressions, again not for every speaker.

4. It might be argued that some other prepositions such as *sous* ‘under’ also show such surface contractions. I concentrate on the two in (9), because they show a clearer phonetic fusion and to avoid overwhelming the reader with data.

5. The variation with *sur* is speaker-dependent. For those speakers who use the latter form, it could be argued that there is no contraction, because the [r] can always be omitted in this word.

6. The use of *un* before vowel initial feminine nouns can be traced back to France, a few centuries ago. Today it is not used by all speakers of QF (Janda, 1998). The speakers who don't use *un* before vowels do not get a different contraction in pre-V position for the feminine and masculine than their pre-C counterpart, that is, they get [syn] or [sən], respectively, if the noun is feminine or masculine.

7. The stratagem is reminiscent of (Zwicky, 1970)'s contrastive stress. Another more general similarity with that article is that its author tried to show how what is often taken to be the obvious analysis is not so.

8. Which leads me to believe that LD is no more a synchronic phonological phenomenon than clitic contraction, just like the “deletion” of the vowel of *la* before a vowel. QF has simply developed a liquid-less shape (in the sense of (Zwicky, 1992)) for *la* and *les* after words ending in a vowel and which cannot be used after a pause, just like in Welsh. The upper class speakers of Santerre et ali., then, were simply avoiding these shapes of the articles and clitic pronouns, but did not avoid the portmanteau prepositions.

9. The only one that was not transcribed was *Comme y a rien de plus plate qu'une vue d'horreur dans une ville fantôme* ‘Since there is nothing more boring than a horror movie in a ghost town’, where *dans une* is pronounced [dyn] (you have to know that *vue* here means ‘movie’ in order to rule out the *d'une* interpretation). As seen in (16), there is good phonological evidence that the contractions involving singular indefinite articles are not lexicalized, at least for most speakers.

10. (24) is a grammatical sentence with a different meaning: the corner store and post office are one entity (some corner stores offer basic postal service). In (18), this special meaning was ruled out because a basement cannot be a wardrobe and vice versa, or if so, I suppose all my speakers were unfamiliar with this situation.

11. There is no noticeable difference between the length of the nasal vowel in P+A and in other non-VF contexts. A certain length can always be observed. This is nicely illustrated by the popular spelling *dans*, for the SF equivalent *dans la*, which speakers cannot distinguish from the non-feminine *dans*. Perhaps when VF was still active, an extra length could be observed.

12. Short high vowels are obligatorily lax in closed syllables.

13. (McLaughlin, 1986:188) concludes that the strategy for adapting borrowings changed sometime after the beginning of the XXth century.

14. These examples, I believe, were first pointed out by (Reighard, 1986), although he comes to a different conclusion, namely that the opposition between the two sets of vowels is one of tenseness and not length.

## References

Auger, Julie. (1995), *Les clitiques pronominaux en français parlé informel : une approche morphologique*, Revue québécoise de linguistique 24, 21-60.

Baronian, Luc V & Stephanie M. Michelet. (2005), *Grammaticalization: Three apparent counter-examples from French*, talk given at the Linguistic Symposium on Romance Languages XXXV, University of Texas Austin.

Bautier-Castaing, Élisabeth. (1977), *Acquisition comparée de la syntaxe du français par des enfants francophones et non francophones*, Études de linguistique appliquée 27, 19-41.

Desjardins, Richard. (1993), *Richard Desjardins au Club Soda*, Montreal: Fukinic.

Dumas, Denis. (1974), *La fusion vocalique en français québécois*, Montréal Working Papers in Linguistics 2, Montreal: McGill University, Université de Montréal, Université du Québec à Montréal.

Dumas, Denis. (1987), *Nos façons de parler: Les prononciations en français québécois*, Sillery: Presses de l'Université du Québec.

Hopper, Paul J. & Elizabeth C. Traugott. (1993), *Grammaticalization*, Cambridge: Cambridge University Press.

Janda, Richard. (1998), *Comments on the Paper by Perlmutter*, in (Lapointe et ali., 1998), 339-359.

Labelle, Guy. (1976), *La langue des enfants de Montréal et de Paris*, Langue française 31, 55-73.

Lapointe, Steven G. Diane K. Brentari & Patrick M. Farrell (eds.), (1998), *Morphology and its Relation to Phonology and Syntax*, Stanford: CSLI Publications.

Larousse. (1964), *Grammaire du français contemporain*, Paris: Larousse.

McLaughlin, Anne. (1986), *Les emprunts à l'anglais et la phonologie des voyelles hautes en français montréalais*, *Revue québécoise de linguistique théorique et appliquée* 5:4, 179-214.

Meillet, Antoine. (1912), *L'évolution des formes grammaticales*, *Scientia (Rivista di Scienza)* 12:26, 6, as reprinted in 1958, in *Linguistique historique et linguistique générale*, Paris: Champion.

Neumann-Holzschuh, Ingrid. (1985), *Le créole de Breaux Bridge, Louisiane : étude morphosyntaxique, textes, vocabulaire*, Hamburg : H. Buske.

Ouellet, Marise et Linda Thibault. (1996), *Durée vocalique en québécois*, in Dolbec, Jean & Marise Ouellet (eds.), *Recherches en phonétique et en phonologie au Québec*, Centre international de recherche en aménagement linguistique.

Paradis, Claude & Denise Deshaies. (1990), *Rules of Stress Assignment in Quebec French: Evidence from Perceptual Data*, *Language Variation and Change* 2, 135-154.

Reighard, John. (1986), *Une analyse concrète du système vocalique du français montréalais*, *Revue québécoise de linguistique théorique et appliquée* 5:4, 281-308.

Santerre, Laurent, Danielle Noiseux & Luc Ostiguy. (1977), *La chute du /l/ dans les articles et les pronoms clitiques en français québécois*, in Paradis, Michel (ed.), *The Fourth LACUS Forum*, Columbia, SC: Hornbeam Press, Inc., 530-538.

Tremblay, Mireille. (1999), *Du statut des prépositions dans la grammaire*, *Revue québécoise de linguistique* 27:2, 167-183.

Walker, Douglas C. (1984), *The Pronunciation of Canadian French*, Ottawa: University of Ottawa Press.

Zwicky, Arnold M. (1970), *The Free-Ride Principle and Two Rules of Complete Assimilation in English*, *Proceedings of the Chicago Linguistic Society* 6, 579-588.

Zwicky, Arnold M. (1992), *Some Choices in the Theory of Morphology*, in Levine, Robert I. (ed.), *Formal Grammar: Theory and Implementation*, London: Oxford University Press, 327-371.

## Chapter 3

# THE A'S AND BE'S OF ENGLISH PREPOSITIONS

Andrew McMichael

*Université de Toulouse 2, CNRS-ERSS - France*

[andrew@enstimac.fr](mailto:andrew@enstimac.fr)

**Abstract**

The focus of this paper will be to show how English prepositions and adverbial particles are genetically related from a morpho-syntactical perspective. To this end, the history of these grammatical items will shed some light on their internal structure and grammatical function. It will be shown that many spatial adverbs are reduced prepositional phrases thus confirming some grammars that equate the PP with an adverb. Based on this morpho-syntactical analysis, we can propose two broad classification systems depending on the desired focus of study, one morphological, and one semantic. The origin and morphology of an adverb/preposition will identify it as being either simplex or compound, with some entailments for identifying its grammatical function. The semantic classification will depend on the origin and core meaning of the prefixed preposition, broadly either association or separation. From a typological perspective, it would seem that other languages display the formation pattern of compounding a preposition with a lexical item to form a new preposition. Lastly, we will try to compare English with how French has grammaticalised some of its spatial relations<sup>1</sup>. Hopefully, the patterns and classifications observed will be useful for computational linguistic systems.

**Keywords:**

Grammaticalisation, spatial adverbs, prepositions, PP constructions, formative patterns, morphology.

### 1. Some definitions

Without going into the long-standing debate on the precise grammatical characterisation of the adverb as opposed to the preposition, we can assume some working definitions for the present purpose:

- 1 Prepositions are spatial grams<sup>2</sup> that perform essentially linking functions to a lexicalised landmark of some nature.
- 2 Adverbial particles (or adverbs) are spatial grams that essentially focus a state without necessarily lexicalising the landmark to which it refers.

To illustrate these definitions adapted from O'Dowd (1998), let us take the gram *down* which answers both, but which can be construed as preposition or adverb depending on the contextual data.

(1) [...] *she found herself falling down a very deep well.*

(2) *She knelt down and looked along the passage [...].*

The landmark of (1) is the *well*, whereas that of (2) has been sublexicalised but we can reasonably suppose that **down** refers to the floor or the ground. However, in (3) the implicit landmark is open to interpretation.

(3) *Take pen and ink and write it down.*

Both **downs** of (2) and (3), however, focus on the state reached after the action just as the **down** of (1) indicates the direction of the goal implied by the landmark *well*, i.e. the bottom. In other words, the event and the participants of the utterance are located relative to a location or a state. As states can be conceptualised as locations (Lakoff & Johnson 1980), the correlation between the landmark of the preposition and the state implied in the adverb is easy to make on semantic grounds. However, what if morphology also corroborated these observations?

## 2. Corpus data

The number of English prepositions and adverbs is between 60 and over 130, the figure varying depending on the definition of the terms. Of the adverbs, Dwight Bolinger (1971:18) estimates that as many as a quarter are used mainly, if not exclusively, in nautical language. In his discussion of the *Phrasal Verb in English*, Bolinger quotes George Meyer's corpus findings (unpublished paper, c. 1969) concerning the 17 most frequently used 'prepositional particles'. These are reproduced here:

(4) **about**, **across**, **along**, **around**, **aside**, **away**, **back**, **by**, **down**, **in**, **off**, **on**, **out**, **over**, **through**, **under**, **up**.

The list can be expanded to include Hill's (1968) frequent particles:

(5) **ahead, alongside, forth, forward, past, together.**

To this list finally can be added Spasov's (1966: 13; 24) frequency counts in Modern English:

(6) **aback, above, after, again, apart, astray, asunder, athwart, before, behind, below, between, round.**

Bolinger's nautical terms include:

(7) **abaft, abeam, aboard, aft, aground, aloft, amidships, aport, ashore, astarboard, astern, overboard.**

One is immediately struck by the number of words in these lists that start with the letter **a-** (27/48), and several commencing with **be-** (4/48). If we do not consider frequency as the defining parameter for devising the list and simply take Hill's 1964 list of prepositions and adverbs, the count is 64 of which 12 start with **a-**, and 8 with **be-**. However, Hill does not include the specifically nautical collection. Whatever the list, the number of A's and BE's, as I will call them, is surprising. The obvious question is why is this so, and what can it tell us about prepositions and adverbs?

A glance in the ODEE<sup>3</sup> will begin to lift a corner of the veil. In order to understand the pattern which is revealed by the preceding observation, let us start with the BE's as they are less numerous and more transparent in their structure. This will also enable us to put a finger on some general features of what turns out to be a grammaticalisation process<sup>4</sup>.

### 3. The general formative principle

Etymologically, BE- comes from **be**, a very ancient preposition going back as far as Indo-European (*\*mbʰi*). In corpus counts, such as in Francis and Kucera (1982), BY is usually realised as a preposition, rarely as an adverb (1% of the time). The preposition occurs typically before nominal items such as:

(8) *be-hind, be-fore, be-side, be-tween* (i.e. two), *be-low* (i.e. lower part).

A lot of these old nouns are hardly recognisable as such today by non-specialists. In the first three cases, we have relatively transparent examples of body-parts having been grammaticalised into prepositions (see Claudi and Heine 1986, Heine, Claudi, and Hünnemeyer 1991, and Heine 1997). The

formative pattern can thus be set out as:

$$(9) p + N \Rightarrow pN$$

in which p stands for a grammaticalising preposition and N a bare noun (in most cases). The syntactical pattern is in no way surprising as it follows the general form of prepositional phrases, e.g. *in (p) the house* (NP). However, the morphology of N draws attention to its high degree of grammaticalisation in that:

- a) - N does not allow a determiner, hence N and not NP,
- b) - N does not allow inflection,
- c) - N has connected to the preposition (p is a quasi-prefix),

and, optionally:

- d) - phonological reduction occurs,

e) - N is no longer morphologically or semantically identifiable as having once been a separate word (e.g. *-tween* < *tweonum*, *tweoh* = *two*). A corollary of this is that those Ns which are morphologically identifiable are not perceived as having a connection with their referent in the real world.

One feature is specific to the preposition:

f) - The preposition has also suffered phonological reduction or aphaeresis and, in some instances, may have altogether disappeared (as will be seen with the A's).

These general features are typical markers of the process that led nouns to be grammaticalised into adverbs by means of a preposition. However, not just nouns are concerned by this formative process. Some verbs and especially adverbs have also combined with a preposition to give a new adverb/preposition. The process was especially productive during the Middle English period, but the pattern is much more ancient<sup>5</sup>. The following discussion will however be limited to p+N constructions.

#### 4. Adverb formation

We are now in a position to conclude that a lot of particles in adverbial function (cf. §:1 definition (2)) can be derived from prepositional phrase structure, i.e. p + NP. The process is that of grammaticalisation of p+N => pN. According to definition (2), the adverb does not necessarily connect to an overt or lexicalised landmark. This can be explained now because the landmark is already contained in the adverbial preposition. The main difference between the two types of landmarks (prepositional and adverbial) is that the second has become highly abstract and is no longer specifically lexicalised as it is with the preposition.

This conclusion does not attempt on universality: all compounds are not adverbs and all simplex forms are not just prepositions<sup>6</sup>. However, the ubiquitous categorical nature of prepositions and adverbs is now clearer as their genetic relationship and structure have been revealed.

It may be added that in order for the adverb to gain prepositional function again another preposition is usually added, such as *of* or *to*, after the gram to link it with a relevant landmark.

#### 5. The A's: a more complicated origin

The prepositions commencing with A- follow the formative pattern outlined above. However, the main difference lies in the origin of the grammaticalising preposition (Gp). Feature f) has become aphaeresis and only an etymological dictionary can supply the original Gp. In order of frequency, ON is the main instrumental preposition:

(10) about (*on-butan*, *on bi-utan*), above (*on-ufan*, also *bi-ufan*, *bufan*), again (*ongeon*, *on gegin*), ahead (*on head*), aside (*on side*), away (*on way*), back (*on baec*); the list of nautical terms was formed mainly with ON.

ON is followed by BY in terms of frequency (cf. §:3). Next, we have IN:

(11) across (*in crosse*, *on croiz*; OF: *en croix*), amidships (LG: *in mid scheeps*) around (OF: *en rond*), into (*in to*; *on to*).

It should be noted that in Old English and in earlier stages of the language, ON and IN were not clearly distinguished semantically as attested in OE and ME texts, and there was a clear preference for ON. Many of the INS were translations from Old or Norman French. Away could also be classified in the INS as the formation was based originally on *en route/chemin*.

OF / OFF (also ultimately related historically to UP and OUT):

(12) after (*aef hinder*), down (*of dune*, i.e. from the hill), over (*ofer, yfer; \*uper-*).

And related to OF, AND-/ANTI -:

(13) along (*and lang*, i.e. *anti* = against, opposite + *lang*).

This preposition is relatively unique in its formative power with only one preposition. However, it may be found in the verb *answer* (*and- swere*: swear or declare back), and, of course, in the conjunction *and*.

Space does not allow a complete list with the origin of all the prepositions. But the demonstration of the principle is clear. Only a few prepositions, that I have called the Grammaticalising Prepositions (see §:8 table 1 below), participated in this formative process.

Although this historical and grammatical account of the formation of prepositions and adverbs (spatial grams) may seem like a digression, it was a necessary one in order to understand what etymology and morpho-syntax can bring to understanding and classifying these grams. Based on these observations, we can propose two classification systems that do not rely solely on syntactic order or semantic considerations.

## 6. Simplex and compound prepositions: a classification

It can now be seen that many, if not all, English spatial grams may be divided between simplex and compound. The simplex category contains usually the most ancient prepositions, e.g. *in, on, off/of, by, with*, and also *out, up, to, at, through*<sup>7</sup>. Of these, only the first five were frequently grammaticalising prepositions. On the other hand, the compound category contains the vast majority of prepositions and adverbs (cf. lists (4), (5) and (6) in §:2). It may be predicted that a simplex gram can function as both preposition and adverb—only syntax can tell them apart. Conversely, compound grams with a substantive as second element in the combination are highly likely to be adverbial in nature often functioning as prepositions only when followed by another preposition and a LM (e.g. *away from X, ahead of X, back to X, etc.*). Of course, this statement is only a generality as grammaticalisation is an ongoing process that tends to blur the line between adverbial and prepositional functions (e.g. *beside, beside X, down, down X*).

## 7. The prefixed Gp as a cognitive functional marker

The class of prepositions as grammatical words is considered to be closed. There are at least two examples which show this is not necessarily so, and that the form of the prefixed element A- could possibly motivate new additions.

The first example is the preposition *amidst* which derived from the simplex gram *mid* which had the meaning of modern *with* and *among*. According to Marie-Line Groussier (1984:1195), the preposition was replaced by *with* in Middle English due to the fact that its uses had lost all reference to space and its meaning had become solely metaphorical. The spatial uses of *mid* in the sense of *among* were taken over by *amidst*. This preposition was formed in the 16c. on the phonological contraction of *on/in the mid(st) of* and obeys the principles of adverbial particle formation set down in §:2.

The second case is more modern but has not left any attested examples yet. Len Talmy in his 1985 paper on “How Language Structures Space” discusses the complex spatial knowledge that comes into play with the use of various prepositions. In a long paragraph, he characterises the spatial dimensions and orientations that are required to understand *across* in “*The plow is across the field.*” (1985:228). Talmy then points out that there is no specific preposition to condense the information needed to represent: “*This field is plowed in concentric circles. Look at the middlemost furrow. There is a pit dug at one point of it. The plow you are looking for is in that pit.*” such as in “*the hypothetical preposition apit in:* “*\*The plow is apit the field.*”” (1985:229). Talmy invents another preposition p. 270 “*aflat*” on the same formative principle of the addition of a grammaticalising prefix A-.

These inventions would tend to show that the Gp is a cognitive reality. The prefixal A- especially, marks off words to indicate their grammatical function. The syntax of the construction is also to be taken into account in the parsing of the utterances.

## 8. Cognitive schemata of grammaticalising prepositions: an alternative categorisation

Another classification schema can be proposed based on the grammaticalising prepositions. More specifically, it can be noticed that the prototypical meanings of the prepositions fall into two or three main categories. The first is the proximity schema, which will be seen to be part of two more general schemata: association and dissociation.

### 8.1 by: the Proximity Schema

Historically, BY can be reconstructed as *\*mb<sup>h</sup>i* in IE. Watkins (1985:3) ultimately derives *\*mb<sup>h</sup>i* from *\*ant-* (meaning front, forehead) which can be

correlated with the *and-* of *along*. In OE, the root gave two words, BY and the defunct *ymbe* replaced in the 16c. by *around*. The meanings of each can be characterised as a relationship between two equidistant entities: BY is relative to a point, whereas *ymbe* refers to a circle<sup>8</sup>. Both indicate proximity with the landmark (LM).

Tr    ⊙ $\iff$ • ) LM

Figure1: PROXIMITY

In figure 1, the LM is perceived either as a line or a point relative to the Trajector (Tr). The trajector is the located entity, which is either moving or potentially movable, or is perceived to be less permanent than the landmark. Tr refers to the entity that is Subject of the verb, or its Object and which is contextualised by the prepositional phrase containing the LM. The proximity schema shows one of the defining features of prepositions, i.e. the interdependency of the trajector and the landmark. The double arrow illustrates the dual direction of this relationship. Schema (1) is however a sub-schema of a more global cognitive schema represented below in (Figs. 2 and 3).

## 8.2    *in, on, of: the Association/Dissociation Schema*

BY and IN, ON and OF are the most typical Gp's. Semantically, ON and IN are associative prepositions that derive historically from a concept of close contact with a landmark. As was pointed out above in §:5, the modern meanings of “contact with the surface of an entity” and “containment within an entity” were not clearly specified in earlier stages of the language.

OF and OFF are dissociating prepositions ultimately related to UP and OUT in IE. The second F on OFF was added in ME to distinguish between purely prepositional uses (OF) and more adverbial functions (OFF). FROM is related to FORE, that is, in relation to the front of an entity, yet separated.

All of these meanings and etymologies point up the most prototypical function of the preposition, that of relating two entities together while specifying their independence as individual entities: at the same time independent yet interdependent. Their relationship will be either one of attraction, or ASSOCIATION, or one of separation or DISSOCIATION. The schemata are laid out in figures 2 and 3. The large arrowheads point in the direction of the main association but there remains an implicit reference back to the entity they contextualise.

$$\text{Tr} \odot \Rightarrow \bullet ) \text{LM}$$

Figure 2: ASSOCIATION

$$\text{Tr} \odot \Leftarrow \bullet ) \text{LM}$$

Figure 3: DISSOCIATION

The grammaticalising prepositions can thus be categorised as in table 3.1 below.

ASSOCIATION	DISSOCIATION
<b>on/in</b>	<b>off/of</b>
<i>ante</i>	<i>by/ymbe</i>
with/mid	<i>with</i>
to/at	for(e)/from
up	out

Table 3.1. Associative and Dissociative grammaticalising prepositions

This table represents only the simplest prepositions that mainly participated in the grammaticalisation process of forming the compound prepositions. The prepositions in bold case have been discussed above. The other prepositions mainly partook in grammaticalising other spatial adverbs or verbs (e.g. *to* in *towards*, *together*, etc.). *By* participates in both schemata. It will come as no surprise that the compounds formed with the items of one of the columns can also be classified as associating or dissociating accordingly.

Although the classification proposed is somewhat simplifying in its approach and results, I would suggest that it globally obeys a general principle of cognitive economy in categorisation. That is to say that the broad schema provided by this classification system can then be fine-tuned to meet the pragmatic realities of the language user. Thus, overlaps and contradictions may appear in actual usage, which are normal occurrences in language variation. To wit, take the example of the historical change of *with* from its original meaning of *against* (\**wi-tero*: separation of two landmarks originating from a central point and comparing their increasing distance) to its modern sense of close association. The ambiguity is manifest in the example *he fought with him* which can

mean he fought either against or alongside the other person.

Another example of a different kind is found with *home*, which Bolinger (1974:18) classified as an adverb. It is clearly on the long road to grammaticalisation but it is arguable whether it is through the intermediary of less productive Gp's, such as *at* or *to*, or whether it is a self-made preposition like *through* was.

## 9. Extensions of the Pattern

The grammaticalisation formula in (9) can now be refined and rewritten thus:

$$(17) p + N \Rightarrow Gp + N \Rightarrow GpN$$

The features set out in §3 concerning GpN, more specifically the first three and the last, can also be found incidentally in a syntactic construction such as the so-called present progressive tense. Historically the construction derived from a subject, the inflected verb *be*, and a nominal form of a verb (V+ING) preceded by the preposition **ON**. For instance:

$$(18) ME: *He waes on huntyng.*$$

The sentences meant literally that “He was **in** the act of hunting”. The old form can still be heard in archaic PDE sentences like “The times, they are a-changing.” which confirm features a), b), c) and f). This hypothesis is dealt with in more detail in Lapaire & McMichael (2001).

The functions of the elements of this formula are also reminiscent of some of the general principles of word formation in English. The prepositional adverb *sideways* might look like it represents an intermediate stage between noun formations of the form N+N (e.g. *sidewalk*) and a formation of (Gp)N+N (e.g. (*on/a*) + *side* + *ways*). A closer look reveals that not only was *side* an adverb, but *ways* was originally *-wise* (meaning: in the manner of) and that it was governed by the Gp **ON**<sup>9</sup>. However, without this specialist knowledge, the layman could construe that there was probably reanalysis and transfer by analogy from one construction to the other resulting in *sideways* and other examples. Whatever the case, such visual parallels as these are evidence again of cognitive economy at play deep in the language with realisations that are not directly identifiable with analogous structures.

## 10. Language Typology

Soteria Svorou (1994:80-6) has many examples of prepositions from other languages<sup>10</sup> that seem to display the same pattern of grammaticalisation. Of

particular interest are some Middle Welsh grams such as *ar ol* (after < *ar* = **on** + *ol* = track), *ar dwrs* (in front of < *ar* + *dwrs* = door), *ar hyt* (along < *ar* + *hyt* = length), *ar ben* (on top of < *ar* + *penn* = head), *ar uchaf* (upon, over < *ar* + *uchaf* = top). The Celtic influence on English is negligible from a lexical point of view but this may be an instance of some syntactical influence. However, the question remains open for the present until further research has been done, including comparison with Scandinavian formative patterns.

Many languages of the world have resorted to the Gp+N formative structure for their prepositions<sup>11</sup>. Further research is required to ascertain whether they have adverbial functions comparable to the English formations as Svorou (1994:51) excluded spatial adverbs occurring only in intransitive constructions from her study.

If we look at how one language has grammaticalised space, French provides some evidence of slightly different formations with analogous results. The French prepositions are very like English ones in that they are free and precede the landmark/NP. Their function is also identical and only differs sometimes in their interpretations. Literal translations are not always possible as the translations into French of examples (1), (2) and (3) from §:1 show.

(19) ...*elle tombait dans un puits très profond.*

The French preposition *dans* means literally *in*, the locative preposition<sup>12</sup>. The idea of *down*, the direction, is derived from the contextualising LM *puits* = *well* and the verb *tomber* = *fall*.

Conversely, the adverbial counterparts of the prepositions were prefixed to the verbs much like Latin and are now not perceived as being additions.

(20) *Elle s'agenouilla et regarda le long du couloir.*

The prefix *a-* comes from Latin *ad-* meaning *at*, i.e. contact between two entities. The verb indicates the action of kneeling and the adverb focuses on the end state of the action with LM (the floor) sublexicalised.

(21) *Prenez plume et encre et consignez-le par écrit.*

This sentence is one translation but possibly not the most frequent equivalent. It has simply been chosen to highlight the adverbial prefix *con-* (Latin: *with* or *against*) absent in other translations. The French language, and other Latin-origin languages, establishes different forms of cognitive connections with the LM but these relationships are still coherent with the schema of ASSOCIATION and the members of that category. *Con-*, *with*, and *down* are members and hence are motivated as conceptual equivalents.

The comparison of the two languages also shows more clearly the adverbial status of the grams but a diachronic study reveals the same broad spatial reference system of contact and dissociation. The main difference between the preposition and the adverb/preverb is the reference to an overt or to an implicit landmark.

## 11. Conclusion

The main purpose of this paper was to show how a study of the morphology of the words commonly referred to as prepositions could help to establish other classification criteria and to reveal the genetic relationship between prepositions and spatial adverbs in English. We can conclude by saying that the apparent polysemy of these words is motivated and a common semantic core can be found hence giving some criteria for a broad classification system. On the other hand, the morphology of these items also points to the importance of taking syntactical criteria into account when defining the grammatical functions of the grams, notably by recognising the importance of the fixed word-order of the prepositional phrase and its various manifestations in the morphology of words. The morphological study shows that the adverb and the prepositional phrase can often be one and the same.

## Notes

1. Conventional abbreviations will be used throughout this paper, namely: N and NP = noun and noun phrase; PP = prepositional phrase; OE, ME, PdE = respectively Old, Middle and Present-day English; IE = Indo-European; LG = Lower German; OF = Old French; an asterisk (\*) before a word marks either a reconstructed form or an incorrect or ungrammatical phrase. Upper case is used for concepts or roots.
2. In the sense of Svorou (1995:31) i.e. Bybee's (1986) short grammatical morphemes with a semantic reference to space.
3. The Oxford Dictionary of English Etymology (1966) edited by C.T. Onions.
4. Grammaticalisation is defined as the historical process by which a lexical item takes on a grammatical function.
5. John Hewson (1992) demonstrates that the prepositional phrase is the oldest fixed syntactic structure in IE.
6. *Up* and *out* are perfect examples of simplex forms which function typically as adverbs. It is interesting to note that neither is used frequently in the grammaticalisation process.
7. *Through* is special in the sense that the noun, meaning "hole", grammaticalised without the means of a grammaticalising preposition. It is also slightly more recent than the others having no IE form but many parallels in daughter languages. All the Gp's are said to have been derived from nouns in IE but their original meanings are difficult to recover. *Through* is also interesting from this respect.
8. Groussier (1984:725;743)
9. ODEE:1009
10. To name only a few of representative diversity: Melanesian Pidgin, Middle Welsh, Persian, Margi, Modern Greek, Indonesian, Basque (although Basque, and other languages not mentioned, have locative suffixes rather than pre-posed or prefixed prepositions).
11. Cf. Svorou 1994: Appendix E for lists of prepositions from 26 languages of the world.
12. It may be noted that the French preposition is itself the result of the combination of *de* and *en*, two other prepositions following the more specific formative pattern p+p.

## References

Bolinger, Dwight (1971) *The Phrasal Verb in English*. Cambridge, Mass.: Harvard University Press.

Claudi, Ulrike and Bernd Heine (1986) "On the metaphorical base of grammar." *Studies in Language* 10: 297-335.

Francis, W. Nelson, and Henry Kuçera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Groussier, Marie-Line (1984) *Le système des prépositions dans la prose du vieil-anglais*. Paris 7, Thèse dirigée par A. Culoli.

Heine, Bernd (1997) *Cognitive Foundations of Grammar*. Oxford: Oxford University Press.

Heine, Bernd, Ulrike Claudi and Friederike Hünnemeyer (1991) *Grammaticalization: A Conceptual Framework*. Chicago: University of Chicago Press.

Hewson, John (1992) "The IE evolution from word-structure to phrase-structure." In *Diachrony within Synchrony: Language History and Cognition*. Günter Kellerman / Michael D. Morissey (Eds.), 1992; Frankfurt am Main: Peter Lang. 395-410.

Hill, L.A. (1968) *Prepositions and Adverbial Particles: An Interim Classification, Semantic, Structural, and Graded*. London: Oxford University Press.

Hopper, Paul J. and Elizabeth Clos Traugott (1993) *Grammaticalization*. Cambridge: Cambridge University Press.

Lakoff, George and Mark Johnson (1980) *Metaphors We Live By*. Chicago: University of Chicago Press.

Lapaire, Jean-Rémi. and Andrew McMichael. (2001), "A new spatial script for the English progressive." *Cognition in Language Use: Selected papers from the 7th International Pragmatics Conference*, Vol. 1. Enikö Nemeth (ed.), Antwerp: International Pragmatics Association. 226-42

McMichael, Andrew (2001), *De l'origine et du développement de la combinaison verbe+ particule en anglais*. Thèse ; Toulouse: Université de Toulouse-Le Mirail.

Meyer, George A. (c. 1969) Unpublished paper on phrasal verbs (quoted by D. Bolinger 1971:17).

ODEE = *The Oxford Dictionary of English Etymology* (1966) Edited by C.T.Onions. London: Oxford University Press.

O'Dowd, Elisabeth M. (1998) *Prepositions and Particles in English: A Discourse-functional Account*. Oxford: Oxford University Press.

Spasov, Dimiter (1966) *English Phrasal Verbs*. Sofia: Naouka i Izkoustvo.

Svorou, Soteria (1994) *The Grammar of Space*. (Typological Studies in Language, 25.) Amsterdam: John Benjamins.

Talmy, Leonard (1983) "How Language Structures Space." In *Spatial Orientation: Theory, Research and Application*. H. Pick and L. Acredolo, eds. New York: Plenum Press, pp 225-282.

Watkins, Calvert (1985) *The American Heritage Dictionary of Indo-European Roots*. Boston: Houghton Mifflin.

## Chapter 4

# TYPOLOGICAL TENDENCIES AND UNIVERSAL GRAMMAR IN THE ACQUISITION OF ADPOSITIONS

David Stringer

*Faculty of Humanities and Social Sciences*

*Mie University, 1515 Kamihama-cho*

*Tsu, Mie 514-8507, Japan*

[stringer@human.mie-u.ac.jp](mailto:stringer@human.mie-u.ac.jp)

**Abstract** Variation in the linguistic expression of trajectories in motion events has been attributed to a widely-accepted binary typology: ‘satellite-framed’ languages (e.g. English, Russian, Chinese) generally encode direction in adpositions or affixes, whilst ‘verb-framed’ languages (e.g. Japanese, French, Arabic) do so in verbs (Talmy, 1985, 1991). However, elicited production data from a first language acquisition experiment comparing directional predication in Japanese, French and English reveals a common syntax of PPs, and no particular language setting, with all variation due to differences in the properties of individual lexical items. I adopt and extend a theoretical framework in which interpretable features interact through syntax, and a layered PP structure is subject to universal principles of interpretation. These findings are in accordance with the hypothesis that the lexicon is the primary locus of crosslinguistic syntactic variation in argument structure.

**Keywords:** acquisition, lexical semantics, motion events, path, PP, predication, preposition, satellite

## 1. Introduction

The syntactic category P (subsuming prepositions, postpositions, and certain affixes considered as incorporated P) plays a greater role in some languages than others in the framing of motion events. According to Talmy’s (1985, 1991, 2000) widely-accepted binary typology, whilst ‘satellite-framed’ languages such as English, Russian and Chinese generally encode direction in P (e.g. ‘swim *across* the river’ / ‘run *into* the cave’), ‘verb-framed’ languages

such as Japanese, French and Arabic typically do so in V (e.g. ‘*cross* the river swimming’ / ‘*textitenter* the cave running’). This crosslinguistic generalization groups together vast numbers of languages independently of cultural traditions or social factors, and points to intriguing limitations on the range of variation in the way motion events are linguistically represented. In order to bring fresh evidence to bear on formal syntactic aspects of this typology, an original first language experiment was devised and conducted with Japanese (J), French (F) and English (E) children and adults. Whilst the results are broadly supportive of the typology, they also highlight weaknesses in the simple binary distinction, and show that despite much lexical variation within and across the three languages, they appear to share a common syntax of motion events. I draw on acquisition data, constructed examples, and generative investigations of PP structure in proposing that children are able to acquire the grammar of directional predication through the combination of two factors: (i) innate knowledge of a universal syntax of PPs; and (ii) the development of a language-particular lexicon, which is able to package grammatically-relevant concepts into lexical entries. This paper is organized as follows. Section 2 presents the satellite-framed / verb-framed distinction, describes the experimentation, and presents elicited production data. In Section 3, a theoretical framework is proposed for comparative analysis in which features such as LOCATION, PATH and PLACE are operational in syntax, and in which V is merged with a layered PP, the whole being subject to universal principles of interpretation. Section 4 summarizes the findings and relates them to lexical parameterization theory. Syntactic variation in motion events is attributed to differences in the properties of individual lexical items, a finding in accordance with the hypothesis that the lexicon plays a pivotal role in crosslinguistic syntactic variation (e.g. Borer 1984, Chomsky, 1995, Emonds 2000, Fukui 1995).

## 2. A monkey, a parrot and a banana

Talmy’s (1985, 1991, 2000)<sup>1</sup> theory of event structure has at its core a small number of universal semantic elements that are mapped to overt linguistic forms such as prepositions and verbs by a process variously referred to as ‘lexicalization’ (McCawley, 1968), ‘incorporation’ (Gruber, 1965), or ‘conflation’ (Talmy, 1972). Such grammatically relevant concepts include Motion, Manner, and Path. Simplifying Talmy’s definitions somewhat, Motion is movement *per se* in the event, Manner is the way in which the entity moves, and Path is the trajectory followed by the moving entity. Examples of canonical ‘satellite-framed’ and ‘verb-framed’ syntax in Japanese, French and English are shown below, with Path predicates underlined<sup>2</sup>

- *Akira wa umi no naka ni odotte haitta.*  
Akira TOP sea GEN inside LocP dancing entered  
'Akira danced into the sea.'
- *Akira est entré dans la mer en dansant.*  
Akira AUX entered in the sea dancing  
'Akira danced into the sea.'
- *Akira danced into the sea.*

In these examples, the features MOTION and MANNER are lexicalized in a matrix or subordinate V (J: *odoru*; F: *danser*; E: *dance*), whilst PATH is lexicalized following the typology, as either V (J: *hairu* 'enter'; F: *entrer* 'enter') or as P (E: *into*).

In an endeavour to shed light on formal principles at work in the syntax of motion events, an experiment was devised and conducted with the participation of 77 monolingual Japanese, French and English child test subjects, divided into five age groups from 3 to 7 years, and 18 adults in corresponding control groups. There were on average five participants in each age group.<sup>3</sup> A picture-story book was designed whose plot is as follows: a monkey sits in a tree-house about to eat his banana; a parrot swoops in, steals the banana, and flies off. The monkey chases the parrot whilst overcoming various obstacles. A specific manner of motion is associated with a specific path on each page of the book: the monkey slides down his tree, runs under a bridge, jumps over a rock, etc. The monkey and the parrot then encounter a lion in a cave, after which the parrot drops the banana. The monkey catches it and hurries home, going through the same motions for a second time, before eating his banana in peace. Test subjects were asked to say what happens on each page. If they did not describe the path followed by the monkey, a series of prompts was followed to elicit appropriate responses. No PATH predicates were used by the experimenter in such prompts. In this way, 1607 relevant examples of PATH predication could be recorded and transcribed for analysis.

As shown in Table 1, the findings support only a broad interpretation of the lexicalization typology: utterances in which adpositions encoded PATH in the absence of an inherent PATH verb<sup>4</sup> were much less common in Japanese and French than in English, accounting for 15.7 percent (68/432) of the Japanese data, 32.2 percent (131/407) of the French data, and 93.4 percent (438/469) of the English data. However, the three languages fell into discrete response categories in all age groups - the mean percentage ranges being 12.5 - 20.0 for Japanese, 25.8 - 39.4 for French, and 92.8 - 94.0 for English. Not only were the responses ranges categorically distinct, but the confidence intervals on the means were non-overlapping (Japanese:  $0.157 \pm 0.034$ ; French:  $0.322 \pm 0.045$ ; English:  $0.934 \pm 0.022$ , using Wald Š approximation). This three-way distinction clearly calls for analysis beyond a simple binary typology.<sup>5</sup>

Age Group	3yrs	4yrs	5yrs	6yrs	7yrs	Mean
Japanese	12.5	14.3	17.4	20.0	12.5	15.7 (68/432)
French	37.5	25.8	27.9	28.2	40.2	29.0 (131/407)
English	93.2	93.2	92.8	93.8	94.0	93.4 (438/469)

Table 1. Percentages of responses with PATH in PP in the absence of an inherent PATH verb.

Exceptions to the dominant pattern in each language resulted in a wide variety of syntactic types, as shown in the following sets of examples, which include test subject codes for reasons of transparency. The codes indicate the language group (J, F, E), then the age group in years (3, 4, 5, etc.), followed by a lower-case letter for position within the age group. Thus J5b is a Japanese 5-year-old, and the second-youngest in the group. The Japanese data included many lexical and syntactic variations on the theme of the monkey rolling down the hill:

- J5b: *korogeochita*  
roll-fell  
'(he) fell and rolled down'
- J7a: *oka no shita ni korogari nagara itta*  
hill GEN bottom LOcP rolling while went  
'(he) went rolling down the hill'
- J6c: *ue kara korogatte iku*  
top from rolling goes  
'he goes rolling from the top'
- J6d: *yama no ue kara korogatta*  
mountain GEN top from rolled  
'he rolled from the top of the mountain'

The expression of PATH in these utterances may be characterized as follows: J5b used a (non-productive) form of lexical compound; J7a produced a complex PP with a noun specifying geometric information (LocN) and a general spatial postposition (LocP), the manner V merging with a durative complementizer, and tense carried by a deictic verb; J6c's utterance has a complex PP with LocN and a directional postposition (PathP), the manner V as an adjunct, and again tense carried by a deictic; and in the utterance of J6c, we find a directional manner V merging with a directional PP (PathPP).

The French data were also highly varied. As the monkey went under the bridge, elicited responses included the following:

- F3c: *il passe par dessous*  
he passes via underneath  
'he goes under'
- F3d: *il passe en dessous*  
he passes LocP underneath  
'he goes under'
- F4a: *il passe par en dessous*  
he passes via LocP underneath  
'he goes under'
- F6d: *il passe sous le pont*  
he passes under the bridge  
'he goes under the bridge'

- F5a: *il court sous le pont*  
he runs under the bridge  
'He runs under the bridge'
- F7e: *il court en dessous le pont*  
he runs LocP underneath the bridge  
'he runs under the bridge'

The first four utterances exemplify an inherent PATH verb merging respectively with (F3c) PathP and LocN; (F3d) LocP and LocN; (F4a) PathP and LocP together followed by LocN; and (F6d) LocP. The last two French examples illustrate a directional manner V merging respectively with (F5a) LocP and (F7e) a complex PP with LocP and LocN. Note that when *par* Švia and *en* ŠLocPŠ are combined, the directional P precedes the locational P, a fact that will prove relevant when the discussion turns to universal aspects of PP hierarchy.

The English data exemplified various means of expressing the monkey's crossing of the river:

- E3b: *he splashes into it and then gets out*
- E3e: *he swims across the river*
- E4c: *swims over to the shore*
- E5b: *he crosses the river*

In the utterance of E3b, we see non-directional manner V with PathP, and the splitting of the complex trajectory; in that of E3e, we find directional manner V with PathP; E4c combined directional manner V with a P modifier and a PathPP; whilst E5b used PathV with a direct object.

In each language and across age groups, directional manner verbs were merged with locative adpositions in directional contexts, e.g.

- E4d: *he runs in the cave*
- E6e: *he climbs on the top of the hill*
- J3e: *ishi ni jumpu shita*  
rock LocP jump did  
'He jumped on the rock'
- J5c: *soto ni nigeta*  
outside LocP fled  
'He fled outside'
- F3a: *il court dans le trou*  
he runs in the hole  
'He runs in the hole'
- F7d: *il nage de l'autre côté*  
he swims LocP the other side  
'He swims across'

However, non-directional manner verbs were never attested with LocP in such contexts in any language. Examples such as E3e: *he jumps in the river* were typical, yet non-directional manner verbs and onomatopoeia invariably merged with overt PathP, e.g. E3b: *he splashes into the river*; J3b: *ishi no ue kara piyon-tte shita* (rock GEN top from whoosh! did 'he whooshed from the top of the rock'). The above examples of elicited production will be shown to provide support for both a strong theory of universal grammar, and a lexicalist approach to argument structure.

### 3. Lexical variation and single syntax

Despite the highly interesting implications of the verb-framed / satellite-framed distinction in terms of stylistics, rhetoric and translation theory (see e.g. Berman and Slobin, 1994; Naigles and Terrazas, 1998; Ohara, 2000; Ozcaliskan, 2002; Slobin, 1996), the designs and conclusions of previous investigations speak to issues of preference rather than possibility, and language use (performance) rather than language knowledge (competence): they do not distinguish what is grammatical from what is ungrammatical.<sup>6</sup> This follows naturally from the way in which the typology was originally conceived. The criterion for classification of a language in the typology is that the language as a whole selects V or P as its most 'characteristic' expression of PATH, i.e. it is a colloquial, frequent, and pervasive speech pattern (Talmy, 1985: 62). The following sections approach the same syntactic variation from a generative perspective.

#### 3.1 The case of congruent features on V and P

I draw on work by Jackendoff (1990) and Emonds (2000) in positing universal semantic features on lexical items which play a role in the syntax of PPs. In particular, verbs may obligatorily select or optionally merge with adpositions carrying the spatial features LOC (LOCATION) (e.g. *under*), which allows either locational or directional interpretation, PLACE (e.g. *within*), which allows only locational interpretation, and PATH (e.g. *to*) which allows only directional interpretation. In the initial set of three examples, Japanese *ni* and French *dans* are both cases of LocP, unlike English *into*, which is a PathP. However, crucially, if in (LocP) is substituted for into (PathP) in the English example *Akira danced into the sea*, the directional interpretation is impossible due to the non-directional manner verb, leaving a strictly locational interpretation: *Akira danced in the sea*. In the Japanese and French examples, the primary predicate (J: *hairu* / F: *entrer* 'enter') specifies a directional interpretation. However, if the non-directional manner verb 'dance' takes the place of the primary predicate, interpretation again is strictly locational, not directional. Thus in French we derive: *Akira a dansé dans la mer* (*Akira danced in the sea* - 'Akira danced in the sea.'), and in Japanese, we get: *Akira wa umi no naka de odotta* (*Akira TOP sea GEN inside LocP danced* - 'Akira danced in the sea.'). Japanese *de* here replaces *ni* for an independent reason (*de* is required by locational adjuncts to activity verbs, rather than stative verbs). Thus interplay between P and V determines whether the interpretation is locational or directional, and if the verbal and adpositional predicates in these sentences have congruent features, locational interpretation is identical in each language.

In colloquial speech, though not in prescriptive grammar, directional interpretation is possible if a directional manner-of-motion verb (e.g. *run*, *jump*, *swim*, *slide*, NOT *\*dance*, *\*twist*) is merged with a locational adposition: in this case the indirect object may be interpreted as a goal in a directional interpretation, as shown below:

- *Akira wa umi no naka ni jampu-shita/hashtitta*<sup>7</sup>  
Akira TOP sea GEN inside LocP jump-did/ran  
'Akira jumped/ran in the sea.'
- *Akira a sauté/couru dans la mer*.  
Akira AUX jumped/ran in the sea 'Akira jumped/ran in the sea.'
- *Akira jumped/ran in the sea*.

This directional reading is a characteristic of colloquial child and adult speech in each language, as confirmed by both adult informants and the elicited production data of test subjects. It appears to be disparaged in the standard varieties, which encode PATH as an inherent feature of either V or P. However, in appropriately colloquial contexts this lexicalization pattern may even be preferred. For example, a French mother who is with her children in the garden as it begins

to rain might naturally produce the first utterance below, but the second would be most improbable:

- *Allez, courons dans la maison!*  
go-2PL run-1PL in the house  
'Come on, let's run in the house!'
- *Allez, entrons dans la maison en courant!*  
go-2PL enter-1PL in the house running  
'Come on, let's enter the house running!'

It should also be noted that although Japanese and French lack an equivalent directional P to English *into*, they do have directional Ps (e.g. J: *kara* 'from', F: *vers* 'towards'), which can form a syntactic structure of the 'satellite-framed' type. Thus in Japanese, French and English, a directional interpretation is possible in all registers if the feature PATH is inherent in either P or V, or both. A directional interpretation is also possible in colloquial varieties if a LocP merges with a certain class of MANNER verbs, a point which is addressed more formally in Section 3.3 below. To return to the first set of examples and the binary typology, the reason that the English example with *into* finds its nearest equivalents in Japanese and French examples with *hairu* 'enter' and *entrer* 'enter' has nothing whatsoever to do with language-particular grammars, but is simply due to the fact that English *into* has no lexical equivalent in Japanese or French.

### 3.2 PATH and PLACE layers in PP

In order to provide a principled account of the interaction between V and P in directional contexts, I maintain that there is a universal layered PP structure, with a higher functional head hosting the PATH feature, and a lower lexical head hosting the PLACE feature. This idea has been adapted by several syntacticians from Jackendoff's (1990) theory of Conceptual Semantics, in which sentences such as *The deer came from behind the tree* and *Barthez went onto the pitch* are assigned the following semantic representations.

- [*Eventi*COME([*ThingDEER*], [*PathFROM*([*PlaceBEHIND*([*ThingTREE*])])])]
- [*Eventi*GO([*ThingBarthez*], [*PathTO*([*PlaceON*([*ThingPITCH*])])])]

There is accumulating evidence that the *[PATH[PLACE]]* configuration is part of syntactic structure. Van Riemsdijk (1990: 236-237) provides convincing evidence of a higher functional layer in German PPs with circumpositions. In cases where there is a (lower) preposition and a (higher) postposition, only the lower lexical P may assign case, may subcategorize the DP, and may impose idiosyncratic selectional restrictions (among other distinctions). This structure is exemplified below in German.

- [*PP,func*[*PP,lex*[*P,lex**hinter*][*DPder Scheune*]][*P,func**hervor*]]  
behind the barn from  
'from behind the barn'

Koopman (2000) draws similar conclusions about Dutch PPs, and makes the pivotal observation that all spatial Ps in the higher functional projection receive a PATH interpretation, whilst those in the lower projection are interpreted as PLACE.<sup>8</sup> Cinque (1999: 138) points out that the same structural hierarchy can be found in English and Italian, with a 'grammatical P' in a lower projection:

- [*PathPFrom*[*PlacePout*[*pof*[*Dpthe darkness*]]])]

- $[_{PathP} Da[_{PlaceP} dietro[_{Pdi[DP noi]}]]]$   
from behind of us  
'from behind us'

The universality of this structure is further supported by the discovery that in languages that express notions of PATH and PLACE in extended spatial case systems, there is a strict hierarchy of PATH, PLACE and 'grammatical' affixes, which exactly mirrors the PP-internal hierarchy (Van Riemsdijk and Huybrechts, 2001; see also Kracht, this volume). In the example from Lezgian below, the oblique stem marker *-re* appears to mirror the role of grammatical Ps such as English *of* and French *de*. The mirror order can be derived either by successive adjunctions or by feature-checking following insertion of the fully inflected lexical item (Chomsky, 1995), as long as the order of checking is the mirror order of the morphological derivation.

- *sew-re-qh-aj*  
bear-of-behind-from  
'from behind the bear'

A further articulation of the PP structure is necessary to include the spatial nouns which proved so ubiquitous in the elicited production data discussed in Section 2. That a bare N projection carrying geometric information is possible within layered PP is supported by independent investigations such as Ayano (2000) and Holmberg (2002).<sup>9</sup> Elements which can appear as bare LocN in English in the absence of an overt D include (in) *front* (of), (on) *top* (of) and in American English only, (in) *back* (of). In Japanese, N rather than P conveys almost all geometric information. Thus concepts of English *in*, *over* and *towards* are expressed with spatial nouns such as *naka* 'inside', *ue* 'top' and *hō* 'direction'. In French, spatial nouns include *haut* 'top', *bas* 'bottom' and *dessous* 'underneath'. An English sentence such as *He jumped from in front of the train* illustrates a fully articulated layered PP structure with spatial N, as shown below.

- $[_{PathPP} from[_{PlacePP} in[_{LocNP} front[_{PPOf[DP the train]}]]]]$

The semantic element TO is usually covert in such contexts (\*?*He jumped to in front of the train*). That there is a covert PathP in syntax in this case is motivated by the fact that as an empty category, it must be locally licensed by strict adjacency to the verb. When moved into a focus position, directional interpretation is impossible e.g.

- *It was in the lake that Bush jumped.* (\*PATH)
- *It was on the pitch that Zidane ran.* (\*PATH)

On the basis of the research evidence summarized in this section, and considering the implications of crosslinguistic syntactic invariance for first language acquisition, I maintain that:

- The internal structure of PP  $[_{PathPP} \alpha[_{PlacePP} \beta[_{LocNP} \gamma[_{P\delta}]]]]$  is universal, and available at all stages of acquisition.

Lexical entries specified as LocP (e.g. *in*, *under*, *on*) are inserted in the lower lexical P, and are interpreted as PLACE in the absence of a functional projection. If the functional PathP is merged, LOC checks the higher functional feature. For example, if the phrase *[on the pitch]* is merged as an adjunct with the verb *run*, it has a [PLACE] interpretation. If it is merged as a complement, there is a functional projection, whose PATH feature is checked by LOC, deriving a directional interpretation.

For posited universal structures, assuming the 'continuity hypothesis', it is to be predicted that they will be present and inviolable at all stages of acquisition. Thus we should never find errors in violation of this layered PP hierarchy, e.g.

- \**The monkey runs in from the cave.*  
(context: from inside the cave)
- \**The monkey climbs top on the hill.*  
(context: on top of the hill)

In 1608 recorded utterances of path predication, such errors were never attested. That this is so in each language and in all age groups lends further support to the notion that these aspects of phrase structure are part of universal grammar.

### 3.3 A principle of PATH interpretation

As noted in Section 2, prescriptive grammar in Japanese, French and English prefers to spell out the feature PATH overtly on V or P or both, whilst natural colloquial speech in each language also allows a PATH interpretation when directional manner V merges with LocP. We can now formalize this latter observation. This possibility is restricted to verbs which may incorporate a functional PathP. In all three languages discussed here, the equivalents of *run* (J: *hashiru*, F: *courir*) may incorporate an empty PathP, whilst the equivalents of *dance* (J: *odoru*, F: *danser*) may not. As shown above in Section 3.1, in Japanese and French colloquial speech, the equivalents of *run* but not *dance* may merge with LocP in directional contexts. The only reason that English *dance* appears to differ in argument structure from its French and Japanese equivalents is that there is contrastive polysemy among English *into / in*, French *dans*, and Japanese *ni*. Just as in French and Japanese, English *dance* requires an overt PathP complement (e.g. *into, to*) in such cases for the VP to convey directed motion.

In order to account for such interpretive variation across verb classes, I propose the following universal principle:

- In a layered PP [*Path PP*  $\alpha$  [*Place PP*  $\beta$ ]]  $\alpha$  may be covert iff the following two conditions obtain:  $\beta$  bears the feature LOC, which checks the higher PATH feature; and the whole is merged with a PATH-incorporating V.

Despite the range of predicate semantics and syntactic structures in all three sets of data, not one of the 1608 utterances violated this principle. As noted at the end of Section 2, all instances of non-directional manner verbs (including onomatopoeia) invariably merged with overt PathP. In cases where  $\alpha$  is covert, it has variable interpretation. This can be seen from the three-way ambiguity in sentences such as E: *Bob ran under the bridge*, or F: *Bob a couru sous le pont* ‘Bob ran under the bridge’, where the LocP may be interpreted as locational (PP in adjunct position), or as directional with either a TO or VIA interpretation.<sup>10</sup> In Japanese, each interpretation is syntactically spelled out, with geometric information encoded in LocN.

- *Bob wa hashi no shita de hashitta.*  
Bob TOP bridge GEN underneath PlaceP ran  
'Bob ran under the bridge.'
- *Bob wa hashi no shita ni hashitta.*  
Bob TOP bridge GEN underneath LocP ran  
'Bob ran under the bridge.'
- *Bob wa hashi no shita o hashitta.*  
Bob TOP bridge GEN underneath ACC ran  
'Bob ran under the bridge.'

In the first example, the postposition *de* is strictly PlaceP. In the second, *ni* is an instance of LocP, which merges with covert PathP to render a TO interpretation (for discussion see Ayano,

2000: 73-75). In the third case, the VIA interpretation is inferred by assigning accusative case to the spatial noun, ensuring 'object affectedness'. This strategy finds parallels in English examples such as *swim the Channel* (ACROSS inferred) and *walk the road to Santiago* (TO END OF inferred). Syntactic structures thus vary according to the inventory of available lexical items, but the claim that interpretable features on lexical items interact with a universal layered PP structure holds in all three languages.

## 4. Conclusion

Typological patterns in motion events have led to much research on the degree to which languages are 'verb-framed' or 'satellite-framed'. However, despite the validity of such generalizations they remain informal tendencies. The comparative syntactic analysis reported here reveals that at least in respect of Japanese, French and English, (i) syntactic variation is singularly due to variation in the lexicon; (ii) no language as a whole selects a 'setting' to frame motion events, i.e. there is no language-particular grammar involved in this variation; and (iii) a fixed hierarchical internal structure of PP and the same interpretive syntactic principles appear to hold in each language. Such findings concur with one trend in universal grammar research that ties parametric variation to the acquisition of the lexicon (e.g. Borer 1984, Chomsky, 1995, Emonds 2000, Fukui, 1995). Aspects of syntax which are universal are plausibly part of the initial state of the language learner, whilst knowledge of language-particular syntax is acquired via positive evidence in the form of lexical items whose contextual properties are revealed through their syntactic environment. The acquisition data gleaned from the elicited production experiment furnish strong support for the contention that the syntactic structure of layered PP and the relevant principles of interpretation are part of the initial state, leaving to children the task of learning their lexicon.

**Acknowledgements** This research was conducted with the support of the Arts and Humanities Research Board, the University of Durham (UK), and a COE research grant from Mie University (Japan). Special thanks to Seiki Ayano, Joseph Emonds, Thierry Guthmann, Anders Holmberg, Bonnie Schwartz and Anne-Sophie Stringer for their advice and constructive criticism.

## Notes

1. Several influential articles by Talmy (e.g. 1985, 1991) have been revised and republished in Talmy (2000).
2. Both constructed examples and verbatim examples from test subjects are given in italics, to distinguish them from glosses and translations. Glosses include the following terms: ACC - accusative; AUX - auxiliary; GEN - genitive; LocP, PlaceP, PathP - Ps carrying the features LOCATION (both locational and directional readings possible), PLACE (only a locational reading possible), or PATH (only a directional reading possible); TE - Japanese TE-form; TOP - topic; e.g. 2PL - second person plural.
3. A fully comprehensive review of this project is provided in Stringer (in preparation). The relation of semantic complexity in P to delays in acquisition is discussed in Stringer (2003).
4. On this analysis, a PATH verb inherently specifies direction in a specific spatial configuration (e.g. *cross*, *descend*, *enter*). NB. There is crosslinguistic variation between supposedly equivalent verbs, which affected the classification. For example, Japanese *noboru* 'climb' and French *grimper* 'climb' must always be used in upward contexts, whilst English *climb* imposes no restrictions on the direction e.g. *He climbed down the cliff, across the ledge, and into the cave*.
5. Japanese and French adults dispreferred the more colloquial PP forms in experimental conditions, the control group percentage means being 3.7 (3/82) for Japanese, 17.9 (21/117) for French and 89.1 (90/101) for English. However, it is noteworthy that in follow-up interviews, the adults judged all the child utterances with PPs to be grammatical in this respect. Thus child-adult differences reveal stylistic preference rather than grammaticality, the latter being the focus of the current investigation.

6. This has always been a key issue in generative grammar: 'The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not the sentences of L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones.' (Chomsky, 1957: 13)

7. The example with *hashiru* 'run' improves with the deictic *iku* 'go'. The bare MANNER verb is unproblematic in colloquial speech, e.g. *Eki ni hashitta* - station LocP ran - 'He ran to the station'. If further spatial information is encoded in the PP it is preferred that direction be spelled out with a deictic verb, e.g. *Eki no naka ni hashitte-itta* - station GEN inside LocP running-went - 'He went running into the station'.

8. I do not assume that the higher functional projection is restricted to PATH.

9. I argue that such N obligatorily lack functional material such as D or plurals (Stringer, in prep.), and as Ayano (2000) points out, bare N e.g. *in front of the train* can be referentially distinguished from DPs e.g. *in the front of the train*.

10. A reviewer questions the possibility of this VIA interpretation with French *sous* 'under'). As noted in Section 2.1, prescriptive varieties prefer that PATH be overt in V or P. Thus there is a clear 'improvement' with VIA lexicalized in *passer* 'pass' e.g. *Bob a passé sous le pont en courant*. However, in colloquial French VIA-*sous* is productively attested, as shown in Section 2 (F5a) and as confirmed by adult informants.

## References

Ayano, S. (2001), *The Layered Internal Structure and External Syntax of PP*. PhD Dissertation, University of Durham.

Berman, R. A., and Slobin, D. I. (1994), *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Borer, H. (1984), *Parametric Syntax: Case Studies in Semitic and Romance Languages*. Dordrecht: Foris.

Chomsky, N. (1957), *Syntactic Structures*. Mouton: The Hague.

Chomsky, N. (1965), *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1995), *The Minimalist Program*. Cambridge, MA: MIT Press.

Cinque, G. (1999), *Adverbs and Functional Heads: A Crosslinguistic Perspective*. Oxford: Oxford University Press.

Emonds, J. E. (2000), *Lexicon and Grammar: The English Syntacticon*. Berlin: Mouton de Gruyter.

Fukui, N. (1995), The Principles-and-Parameters approach: A comparative syntax of English and Japanese. In: M. Shibatani and T. Bynon (eds.), *Approaches to Language Typology*, 327-372. Oxford: Oxford University Press.

Holmberg, A. (2002), Prepositions and PPs in Zina Kotoko. In B.K. Schmidt, D. Odden and A. Holmberg (eds.), *Some Aspects of the Grammar of Zina Kotoko*. Munich: Lincom Europa.

Jackendoff, R. (1990), *Semantic Structures*. Cambridge: MIT Press.

Koopman, H. (2000), Prepositions, postpositions, circumpositions and particles: The structure of Dutch PPs. In H. Koopman, *The Syntax of Specifiers and Heads: Collected Essays of Hilda. J. Koopman*, 204-260. London: Routledge.

Naigles, L. R., and Terrazas, P. (1998), *Motion verb generalizations in English and Spanish: Influences of language and syntax*. Psychological Science 9: 363-369.

Ozçaliskan, S. (2002), *Metaphors we move by: A crosslinguistic-developmental analysis of metaphorical motion events in English and Turkish*. PhD dissertation, University of California, Berkeley.

Ohara, K. H. (2000), Cognitive and structural constraints on motion descriptions: Observations from Japanese and English. *Proceedings of the 2nd International Conference on Cognitive Science and the 16th annual meeting of the Japanese Cognitive Science Society*: 994-997.

Riemsdijk, H. van (1990), Functional prepositions. In H. Pinkster and I. Genée (eds.), *Unity in Diversity*. Dordrecht: Foris.

Riemsdijk, H. van (1996), Categorial feature magnetism: The extension and distribution of projections. Unpublished ms., Tilburg University.

Riemsdijk, H. van and Huybrechts, R. (2001). Location and Locality. In M. van Oostendorp and E. Anagnostopoulou (eds.), *Progress in Grammar: Articles at the 20th Anniversary of the Comparison of Grammatical Models Group in Tilburg*. Amsterdam: Rocquade.

Slobin, D.I. (1996), Two ways to travel: Verbs of motion in English and Spanish, In M. Shibatani and S.C.A. Thompson (eds.), *Grammatical Constructions: Their Form and Meaning*, 195-219. Oxford: Oxford University Press.

Slobin, D.I. (2003), The many ways to search for a frog: Linguistic typology and the expression of motion events. In S. Strömqvist and L. Verhoeven (eds.), *Relating events in narrative: Typological and contextual perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Stringer, D. (2003), Splitting the conceptual atom: Acquisitional evidence for semantic decomposition. *Durham Working Papers in Linguistics* 9: 81-94.

Stringer, D. (in preparation), *Paths in First Language Acquisition: Motion through Space in English, French and Japanese*. PhD dissertation, University of Durham.

Talmy, L. (1985), Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (ed.), *Language Typology and Syntactic Description, Vol.3: Grammatical Categories and the Lexicon*, 57-149. Cambridge: Cambridge University Press.

Talmy, L. (1991). Paths to realization: A typology of event conflation. *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society*: 480-519.

Talmy, L. (2000), *Toward a Cognitive Semantics, Vol. II: Concept Structuring Systems*. Cambridge, MA: MIT Press.

## Chapter 5

# MULTILINGUAL INVENTORY OF INTERPRETATIONS FOR POSTPOSITIONS AND PREPOSITIONS

*An application to Basque, English and Spanish*

Mikel Lersundi and Eneko Agirre

*University of the Basque Country*

*649 pk. - 20.080 Donostia*

*The Basque Country*

{jibleaym, eneko} @si.ehu.es

**Abstract** This article describes a common inventory of interpretations for postpositions (in Basque) and prepositions (in English and Spanish). The inventory is a flat list of tags, based mainly on thematic roles. Using the same inventory for all languages allows to know, for each postposition or preposition, which are the translations under each possible interpretation. We think this resource will be useful for studies on machine translation, but also on lexical acquisition experiments on the syntax-semantic interface that make use of multilingual data (see, for instance (Agirre and Lersundi, 2002)). The method to derive the inventory and the list of interpretations for Basque postpositions and Spanish and English prepositions has tried to be systematic, and is based on (Aldezabal, 2004) and (Dorr, 1993).

**Keywords:** Multilinguality, postpositions, thematic roles, Lexical Conceptual Structures

## Introduction

This article describes an inventory of interpretations for postpositions (in Basque) and prepositions (in English and Spanish). Basque is an agglutinative language, and its postpositions are more or less equivalent to prepositions, but they are also used to mark syntactic functions such as the subject and objects of verbs.

Literature on Basque suffixation phenomena has not agreed yet on a common definition of *postposition*. For some, they are separate from grammatical

cases (absolutive, ergative and dative), and postpositions refer to the rest of cases (instrumental, adlative, possession-genitive, etc.). For others, postpositions are used for combinations of words like “-en parean” (in front of) including the case suffix (-en in this example) and the postposition proper (parean – front of). We prefer to take a simpler and more general view, and refer as postpositions to all cases (grammatical or not) and also to complex postpositions (like “-en parean”). This allows for an integration of all of them in a unified framework.

Our inventory includes interpretations for both arguments and modifiers, in a generic form. As the list of interpretations is common for all languages, a by-product is that it is possible to know which are the possible translations for a given postposition or preposition into the other languages. Table 5.2 shows some possible translation of the Basque instrumental postposition -z into English prepositions.

The table of interpretations is a generic knowledge resource that helped in the acquisition of complex multilingual structure in the framework of the MEANING project (Rigau et. al., 2002)<sup>1</sup>. For instance, (Agirre and Lersundi, 2002) describe a method based on such a multilingual table that links the syntactic function of an argument or adjunct to the semantic interpretation of the argument or adjunct. The method has proved to be effective to disambiguate the occurrences of the Basque postposition -z (instrumental case) in dictionary definitions, using parallel Spanish and English definitions. We are currently applying the method to all postpositions using the multilingual table described in this paper.

Our inventory of interpretations is based on (Aldezabal, 2004) and (Dorr, 1993). Our goal is to deliver a flat list of interpretations in the form of tags. The tags are derived mainly from thematic role tags, but also cover adjuncts and other phenomena. In order to have a common inventory of interpretations, we have to fix first which are the interpretations that we are interested in. This is not an easy task, and we decided to fix the inventory as we were building the table of interpretations.

This article is organized as follows. Section 5.1 presents previous work, followed by Section 5.2, which presents the method used to build the table and the inventory of interpretations. Section 5.3 illustrates the method with an in-depth study of the Basque instrumental case (-z). Section 5.4 presents the analysis of the results. Section 5.5 reviews some remaining problems, and, finally, Section 5.6 draws the conclusions.

## 1. Previous work

As we can read in (EAGLES, 1998), semantic relations were introduced in generative grammar during the mid-1960s and early 1970s (Fillmore, 1968;

Jackendoff, 1972; Gruber, 1967) as a way of classifying the arguments of natural language predicates into a closed set of participant types which were thought to have a special status in grammar. A list of the most popular roles and the properties usually associated with them (adapted from (Dowty, 1989)) is given below.

**Agent:** A participant which the meaning of the verb specifies as doing or causing something, possibly intentionally. Examples: subjects of *kill, eat, hit, smash, kick, watch*.

**Patient:** A participant which the verb characterizes as having something happen to it, and as being affected by what happens to it. Examples: objects of *kill, eat, smash* but not those of *watch, hear, love*.

**Experiencer:** A participant who is characterized as aware of something. Examples: subject of *love*, object of *annoy*.

**Theme:** A participant which is characterized as changing its position or condition, or as being in a state or position. Examples: objects of *give, hand*, subjects of *walk, die*.

**Location:** The thematic role associated with the NP expressing the location in a sentence with a verb of location. Examples: subjects of *keep, own, retain, know*, locative PPs.

**Source:** Object from which motion proceeds. Examples: subjects of *buy, promise*, objects of *deprive, free, cure*.

**Goal:** Object to which motion proceeds. Examples: subject of *receive, buy*, dative objects of *tell, give*.

In linguistic theory, thematic roles have traditionally been regarded as determinant in expressing generalizations about the syntactic realization of the predicate arguments (see (EAGLES, 1996)). In many cases, the interpretation of the prepositions is linked to thematic roles.

(Aldezabal, 2004) presents an in-depth study of 100 Basque verbs, including their argument structure and also mentioning the semantic interpretation of elements, which are related to thematic roles. We have used the link between the argument structure and the semantic interpretation in order to extract possible interpretations for postpositions<sup>2</sup>. This list of postpositions and their interpretation is the main source of our inventory for Basque. Nevertheless, it has some shortcomings:

- Aldezabal's work focuses mainly on arguments, and her inventory of interpretations may miss adjuncts. To get over this gap, we checked

the inventory of interpretations from (Dorr, 1993) which does include adjuncts.

- For a given postposition, some interpretations might be missing. Aldezabal works only on the interpretations that arise during her study of the 100 verbs, and it could be the case that some interpretations for a given postposition do not appear in her data. We tried to cover those missing interpretations with bilingual dictionaries, and the interpretations on other languages.
- Some postpositions might be missing. We will take missing suffixes from Basque grammars. We currently do not include complex postpositions (which are comparable to complex prepositions in English, e.g. the already seen “*-en parean*”, which is equivalent to “in front of”).

Regarding English and Spanish prepositions, our main source is the Lexical Conceptual Structure (LCS) description in (Dorr, 1993; Dorr and Habash, 2002). From the LCS we extract the thematic roles assigned to prepositions, either directly from the description of the prepositions or indirectly from the LCS describing verbs. We also got a table from (Habash, 2002) where each English preposition has a list of possible thematic roles. The same way that we use the term ‘postposition’ for all cases in Basque (grammatical or not), we decided to adopt a uniform representation for prepositions and syntactic functions in other languages as well. For example, we added the dummy  $\phi$  preposition for subject and object positions in English and Spanish. This does not imply any linguistic claim, and is just a practical issue.

Extracting interpretations for all prepositions from the LCS is not straightforward, as the interpretation of some prepositions are not always described in terms of thematic roles (e.g. it might refer to a primitive). For the sake of this paper, when we refer to the LCS of a preposition, we really mean either the thematic role or the primitive that identifies the interpretation of the preposition. Besides, some interpretations for prepositions might be missing, specially for adjuncts. To our experience, the quality and coverage for English prepositions is very good, but the Spanish prepositions are not so well represented.

## 2. Method to obtain the inventory and the multilingual table

First, we have to decide which kind of interpretation we will use in the description of postpositions, and use the same interpretation inventory for English and Spanish prepositions. In order to get this interpretation inventory we had two main sources:

- Izaskun Aldezabal’s semantic interpretation for some Basque postpositions (Aldezabal, 2004).

- Bonnie Dorr's LCS for English and Spanish prepositions (Dorr, 1993).

In addition, we also have used the examples from bilingual dictionaries:

- Bilingual English/Basque and Spanish/Basque dictionaries (Morris, 1998; Elhuyar, 1996).

A first approach to get the inventory list would be to map both Aldezabal's and Dorr's interpretations, and perhaps choose an inventory which is a combination of both. We compared both lists, and realized that sometimes Dorr's interpretations are more specific than those from Aldezabal (this is the case of *perc-perceived item-*); but, overall, Aldezabal's interpretations are more specific than Dorr's (we have *touched theme*, *displaced theme*, etc. instead of a single *theme*).

There is another disagreement between what Aldezabal considers semantic interpretation and Dorr considers thematic role. For example Aldezabal considers as semantic interpretations both *cause* and *path*, and in the LCS representation done by Dorr, these appear as primitives and types. We also realized that *manner* is listed among Dorr's thematic roles, but it is not linked to any preposition in her lexicon.

The problem is that it is very difficult to match interpretations without studying the examples to which they apply. This is specially the case when the interpretations have been given for different languages.

As a method to fix the inventory of interpretations and build the multilingual table, we start on Basque and jump into the other languages via a set of manually tagged bilingual examples from a bilingual dictionary. Previously, we decided to group some of Aldezabal's semantic interpretations (which are too granular) into a single interpretation. After this, the postpositions in the Basque examples are tagged using our interpretations and the tag is copied to the corresponding example in Spanish and English. Finally, we compare the interpretations of Spanish and English prepositions thus obtained with the thematic roles given by Dorr.

This is the method step by step for each postposition:

- 1 Take a postposition.
- 2 Extract examples for this suffix from the bilingual dictionaries. This way we will translations in context of the suffix into the other two languages.
- 3 Look for interpretations of this postposition in (Aldezabal, 2004).
- 4 Study the interpretations, and, when we think interpretations are too fine-grained, join them, controlling that it is coherent with the other postpositions.

5 Tag the Basque examples with the interpretations, and take special care in possible gaps:

- Study if there is any interpretation in the Basque examples that is missing from Aldezabal's list.
- Find Basque examples and English translation for the interpretations that do not appear in the examples from the bilingual dictionary, but which do appear in Aldezabal's list.

6 At this stage we already have a list of interpretations for the Basque suffix, a list of examples for each interpretation, and a list of English and Spanish translations for each interpretation.

7 Each English preposition in the bilingual examples is assigned the Basque interpretation. This is compared with Dorr's LCS for that preposition<sup>3</sup>.

8 At this stage we produce a list of 4-tuples: (postposition - Aldezabal's interpretation - preposition - Dorr's LCS). From the study of the 4-tuples we derive the following:

- A study of the mismatches between both interpretations, including gaps in the interpretation of English and Spanish prepositions, accompanied by a qualitative and quantitative analysis of the mismatches.
- A unified interpretation tag that tries to solve the mismatches, based on Aldezabal's and Dorr's tags, thus yielding a list of triples: (postposition, unified interpretation, preposition).
- A mapping from the unified interpretation to Aldezabal's and Dorr's inventories.

After applying this method to all postpositions, we get a unified inventory of interpretations that is applied to Basque postpositions and English and Spanish prepositions. We also get a mapping between our unified inventory and Dorr's and Aldezabal's inventory.

### 3. Case study with the Basque instrumental postposition

We will illustrate the methodology of our study using the instrumental postposition. First we look for interpretations of this postposition in (Aldezabal, 2004). It is important to take into account that the goal of Aldezabal's PhD work is not the study of thematic roles. She determines the argument structure of some verbs, and arranges them into groups according to their syntactic behavior. During her study she mentions some semantic interpretations of Basque

Interpretation	English-prepositions	Spanish-prepositions
Cause	because of, due to, for, from, in, of, on account of, out of	Φ, a causa de, con, de, por
Content	in, of, with	con, de
Instrument	Φ, by, for, in, on, with	Φ, a, con, en, por
Manner	Φ, at, by, in, on, with	a, de, en
Path	Φ, along, by, by way of, on, round, through	a traves de, por, por delante de, por encima de, sobre
Theme	Φ, about, at, for, in, of, on	Φ, a, acerca de, con, de, en, sobre
Time	Φ, at, by, during, for, in, on	Φ, a, de, durante, en

Table 5.1. The first column shows the interpretations for the Basque instrumental postposition. The second and third columns show the list of prepositions in English and Spanish with a common interpretation after applying the method to all Basque postpositions

Interpretation	Basque example	English translation
Cause	<b>beldurrez</b> isildu ziren	they shut up <b>out of fear</b>
Content	ontzia <b>urez</b> bete zuen	she filled the container <b>with water</b>
Instrument	hirira <b>autobusez</b> joan zen	she went to the city <b>by bus</b>
Manner	eskuz idatzi zuen	she wrote it <b>by hand</b>
Path	lehorrez joan zen	she went <b>by land</b>
Theme	<b>zutaz</b> asko daki	he knows a lot <b>about you</b>
Time	<b>hiru urtez</b> egon ziren han	they were there <b>for three years</b>

Table 5.2. Interpretations for the Basque instrumental postposition

postpositions, but the goal is not to produce an exhaustive list of semantic interpretations.

Sometimes the interpretations she gives to postpositions are very granular, and we have tried to do a list with more general interpretations, joining some of her interpretations. Aldezabal gives 12 interpretations to the instrumental and we joined them into 6 (step 4 above).

After this, we tag the examples extracted from bilingual dictionaries (61 examples) and we check them in order to see if there is any new interpretation for the postposition (step 5). In the case of the instrumental we found a new interpretation. Table 5.2 shows the 7 interpretations for the Basque instrumental postposition.

Once we have tagged the examples extracted from the bilingual dictionaries with the interpretations (step 6), we obtain the equivalences in Table 5.1.

Once we have the database of triples, we compare the interpretations for English and Spanish prepositions obtained so far with the ones we have from Dorr's work (step 7). During this comparison, we will be able to map Dorr's LCS with the ones we have; and, at the same time we will build the 4-tuples

	English	%	Spanish	%
4-tuples	44		33	
Good map	19	43.18	10	30.30
Primitive-role problem	14	31.82	7	21.21
Missing interpretation	10	22.73	7	21.21
Missing preposition	1	2.27	9	27.27

Table 5.3. Evaluation of the mapping of the instrumental postposition with Dorr's LCS. 4-tuples refer to (postposition - Aldezabal's semantic interpretation - preposition - Dorr's LCS)

mentioned on step 8 of the method (postposition - Aldezabal's semantic interpretation - preposition - Dorr's LCS). After building the list, we are able to evaluate the quality of the mapping between the unified interpretations and Dorr's interpretations (Table 5.3).

There are 44 4-tuples between English prepositions and the Basque instrumental postposition, depending on the interpretation, and 33 between Spanish prepositions and the Basque instrumental. From these, we have 19 good links with English, and 10 with Spanish. We say the link is good when our interpretation agrees with a thematic role in Dorr's LCS. Table 5.4 shows some of the mappings occurring in the 4-tuples for the instrumental case. The primitive-role problem line in Table 5.3 relates to the case when Dorr represents what we call an "interpretation" with an LCS primitive. In the case of the instrumental all the mismatches are caused by primitives *cause* (7 for English and 4 for Spanish) and *path* (7 for English and 3 for Spanish). Table 5.5 shows the relevant mapping .

Regarding missing interpretations in Table 5.3, they are mainly caused by *manner* and *time*. Dorr's representation takes *manner* as a thematic role (they also have it as a type), but they have not assigned it to any preposition. This may be because *manner* is not usually part of an argument, and their job focuses on arguments of verbs. We have counted this as a "missing interpretation", and amounts to 6 (English) and 3 (Spanish) of the missing interpretations. Something similar happens with *time*: 3 of the missing interpretations in English ( $\emptyset$ , by, in), as well as 3 of the missing interpretations in Spanish (a, de, en) are caused by *time*. Incidentally the *instrumental* interpretation is missing from English (in) and Spanish (a).

The missing prepositions for English is "on account of". For Spanish, " $\emptyset$ "<sup>4</sup> (4 links), "a causa de", "acerca de", "durante", "por delante de", and "por encima de" are missing.

The process is repeated for all postpositions (see section 5). At this point the final unified interpretations are fixed. After adding the information for all English (and Spanish) prepositions from Dorr (via mapping), the table for the

<i>Unified interpretations</i>	<i>Dorr's th-roles in the LCS</i>
theme <sup>5</sup>	perc
theme	th
theme	info
instrument	ins
content	poss
time	time

Table 5.4. Mapping with Dorr's thematic roles in the LCS

<i>Unified interpretations</i>	<i>Dorr's primitives</i>
cause	cause
path	path (TO, TOWARD, VIA)

Table 5.5. Mapping with Dorr's primitives in the LCS

instrumental is as shown in Table 5.6. Note that the list of Spanish prepositions is shorter, due to the fact that less prepositions were covered in Dorr's work.

#### 4. Overall results

After analyzing all Basque postpositions, their intersection with English and Spanish prepositions, and the comparison with Dorr's LCS, we get the quantitative results as shown in Table 5.7.

Regarding English, most of the mappings are correct. The percentage of missing interpretations is quite high, but most of them are caused by the *manner* and *time* interpretations not being present in the English data (30 and 9 times respectively). Regarding Spanish, *manner* is also missing, but the main problem for Spanish is the lack of coverage of prepositions.

Once we have applied the method to all Basque postpositions, we have built the mapping between Dorr's LCS and the unified list of interpretations. For instance, Table 5.6 shows the definitive list of triples for the instrumental postposition, and the complete set of interpretations and equivalencies between postpositions and prepositions is accessible on the Internet<sup>6</sup>. Table 5.8 shows the main figures in relation with the number of postpositions and prepositions we have used, and the number of triples that we get for each pair of languages.

#### 5. Remaining problems

Once we have built the mapping between Dorr's LCS and the our own inventory of interpretations, there may be some thematic roles without mapping. One example is Dorr's *purpose* thematic role. We have decided to exclude it

<i>Instrumental</i>	<i>English prepositions</i>	<i>Spanish prepositions</i>
<b>Cause</b>	because of, due to, for, from, in, of, on account of, out of	φ, a, a causa de, con, de, por
<b>Content</b>	φ, about, between, by, for, from, in, of, on, out of, with	a, con, contra, de, en, encima de, por
<b>Instrument</b>	φ as, by, for, from, in, of, on, out of, with, without	φ, a, con, de, en, por
<b>Manner</b>	φ, at, by, in, on, with	a, de, en
<b>Path</b>	φ, along, by, by way of, on, round, through	a traves de, por, por delante de, por encima de, sobre
<b>Theme</b>	φ, about, after, against, around, at, before, for, from, in, into, of, on, over, that, through, to, with	φ, a, acerca de, ante, con, contra, de, en, por, que, sobre
<b>Time</b>	φ, about, after, ahead of, around, as, as of, at, back to, before, behind, between, beyond, by, close to, during, following, for, from, in, in relation to, near, on, per, previous to, prior to, pursuant to, related to, relative to, round, since, through, throughout, till, to, until, with respect to, within	φ, a, de, durante, en

Table 5.6. Unified interpretation of the instrumental postposition together with its equivalent English and Spanish prepositions

	<i>English</i>	<i>%</i>	<i>Spanish</i>	<i>%</i>
4-tuples	272	100	207	100
Good map	161	59.19	81	39.13
Primitive-role problem	51	18.75	26	12.56
Missing interpretation	51	18.75	43	20.77
Missing preposition	9	3.31	57	27.54

Table 5.7. Overall evaluation of the 4-tuples (postposition - Aldezabal's semantic interpretation - preposition - Dorr's LCS)

number of postpositions	14
number of English prepositions	123
number of Spanish prepositions	25
number of Basque-English triples	946
number of Basque-Spanish triples	339
number of English-Spanish triples	2796

Table 5.8. Main figures for the whole database

for the time being, because in most of the cases this is a Verb-Verb relation, and we have focused on Verb-Noun relations.

includes

Regarding the amount of prepositions, the list of prepositions in English is quite comprehensive. According to our analysis we think that most of their interpretations are covered, as we have completed Dorr's LCS with those which appeared in the bilingual examples. The situation is worse for Spanish, as we have a lot of missing interpretations and prepositions. Further work is needed in order to get a satisfactory status for Spanish. Regarding Basque we are very satisfied with the coverage, but we need to extend our work to complex Basque postpositions which were not included in this study.

## 6. Conclusions and future work

We have produced an inventory of interpretations that has been used to describe Basque postpositions and English and Spanish prepositions (see Table 5.9). The whole database is freely available in the Internet<sup>6</sup>. Using a single inventory allows to know for each postposition or preposition which are its equivalents on the same language, as well as which are the translations for each possible interpretation. We think this resource will be useful for studies on machine translation, but also on lexical acquisition on the syntax-semantic interface which makes use of multilingual data.

The source of the unified inventory of interpretations has been Aldezabal's semantic interpretations (Aldezabal, 2004) and Dorr's lexicon of LCS for verbs and prepositions (Dorr, 1993; Dorr & Habash, 2002; Habash 2002). We provide a mapping from our inventory to both of them. Their work also provides the main source of interpretations for each postposition and preposition. We have to note that our interpretations try to cover all possible meanings of a preposition when acting as an argument or adjunct of a verb. Dorr's work is relevant because although her description focuses on argument structure of lexical verbs, she also gives importance to adjuncts. She has also analyzed a list of prepositions (including complex prepositions), and once we got the relation between her LCS and the our of interpretations, we have been able to use all the English and Spanish prepositions she has studied. We have to note that we have simplified the LCS: we take an atomic tag, either a thematic role or a primitive, which summarizes the interpretation of the preposition in her LCS.

The method to derive the inventory and the list of interpretations for Basque postpositions and Spanish and English prepositions tried to be systematic. We first extracted the interpretation for Basque postpositions from Aldezabal's work on verbs. We complemented this data with examples from bilingual dictionaries (Basque/Spanish and Basque/English), which also provide English translations. Checking Aldezabal's semantic interpretations for each bilingual

<i>Basque name</i>	<i>English name</i>
agentea	agent
baliabidea	instrument
bidea	path
denbora	time
edukia	content
esperimentatzaire	experiencer
ezaugarria	attribute
gaia	theme
hasierako denbora-kokapena	source-time
hasierako leku-kokapena	source-location
helburua	goal
helburuko denbora-kokapena	goal-time
helburuko leku-kokapena	goal-location
iturria	source
jarduera	activity
kausa	cause
konpainia	company
leku	location
modua	manner
noranzko	direction

Table 5.9. Unified inventory of interpretations

example against Dorr's LCS allowed us to construct a systematic mapping. The main advantage of this method is that we are able to map different inventories of interpretations based on actual examples, rather than the sole intuition of the linguist. The results of this analysis are a database of triples (Basque postposition – interpretation – English or Spanish preposition) plus mappings between our interpretations and Dorr's and Aldezabal's interpretations.

Regarding future work, it is important to remark that the inventory of interpretations and the database is not in a final stage. Some further research needs to be done for a number of issues. Nevertheless, the use of three different sources and the work done extracting the relationship between them gives a strong base to this approach.

More specifically, we would like to find a better treatment of the dummy “ $\phi$ ” preposition, and specify whether there is a “subject” or an “object” relation. We will also need to go beyond verb-noun relationships, and cover all syntactic functions intermediated by prepositions or postpositions.

Regarding Basque we have to incorporate all complex postpositions with their interpretations. Spanish is without doubt the language with worse coverage: we have only 3.31% missing prepositions for English, while 27.54% are missing for Spanish.

## Acknowledgments

Part of the work was carried out during the stay of Mikel Lersundi in the University of Maryland. We want to thank the people in the CLIP laboratory there, and in particular Nizar Habash and Bonnie Dorr. We also want to thank Izaskun Aldezabal for sharing the material on her Ph.D. work before its publication. The work is partially funded by the European Commission (MEANING project, IST-2001-34460), and MCYT (HERMES project, TIC-2000-0335-C03-03).

## Notes

1. <http://www.lsi.upc.es/~nlp/meaning/meaning.html>
2. As mentioned in the introduction, the syntactic function of Basque arguments is marked in the surface by postpositions, that is, each argument has a postposition that determines (ambiguously) the syntactic function of the argument.
3. Remind that we examine not only the LCS of the preposition, but also the LCS of all verbs which subcategorize the preposition
4. For the sake of this article, “ $\emptyset$ ” corresponds to noun phrases without prepositions. In the future, we plan to split “ $\emptyset$ ” into “subject” and “object” syntactic functions.
5. It is important to remark that our theme interpretation has always a *perc* interpretation between English and Spanish prepositions. This happens in the case of the instrumental postposition.
6. <http://ixa.si.ehu.es/Ixa/local/casesuffixes>

## References

Agirre, E. and M. Lersundi (2002). “A multilingual approach to disambiguate prepositions and case suffixes”, *Proceedings of the Word Sense Disambiguation: Recent Successes and Future Directions Workshop*, University of Pennsylvania, Philadelphia, USA.

Aldezabal, I. (2004), *Aditzaren azpikategorizazioaren azterketa aplikazio konputazionalari begira (Analyzing verbal subcategorization aimed at its computational application)*, PhD thesis, University of the Basque Country, Donostia, Basque Country.

Dorr, B. (1993), *Machine Translation: A View from the Lexicon*, Cambridge, Massachusetts, MIT Press.

Dorr, Bonnie and Nizar Habash (2002), “Interlingua Approximation: A Generation-Heavy Approach”, *AMTA-2002 Interlingua Reliability Workshop*, Tiburon, California, USA.

Dowty, D. (1989), “On the Semantic Content of the Notion of Thematic Role”, in G. Cherchia, B. Partee, R. Turner (eds), *Properties, Types and meaning*, Kluwer.

Selection”,

EAGLES (1996), “Preliminary Recommendations on Subcategorisation”. Available in <http://www.ilc.cnr.it/EAGLES96/synlex/node63.html>.

EAGLES lexicon interest group (1998), “Preliminary Recommendations on Semantic Encoding”. Available in <http://www.ilc.cnr.it/EAGLES96/rep2/rep2.html>.

Elhuyar (1996), *Elhuyar Hiztegia*, Elhuyar K.E., Usurbil, Basque Country.

Euskaltzaindia (1985), *Euskal Gramatika Lehen Urratsak-I (EGLU-I)*, Euskaltzaindia, Bilbao, Basque Country.

Fillmore, C. (1968), “The Case for Case”, in E. Bach and R.T. Hams (eds.) *Universals in Linguistic Theory*, Holt, Rinehart and Winston, New York, USA.

Gruber, J. (1967), *Studies in Lexical Relations*, MIT doctoral dissertation. Also in *Lexical Structures in Syntax and Semantics*, North Holland (1976).

Habash, Nizar (2002), “Generation-Heavy Hybrid Machine Translation”, Proceedings of *INLG-02*, New York, USA.

Jackendoff, R. (1972), *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge, Mass., USA.

Jackendoff, R. S. (1990) Semantic Structures, MIT Press, Cambridge, Mass., USA.

Morris, M. (1998), *Morris Student dictionary*, Klaudio Harluxet Fundazioa, Donostia, Basque Country.

Rigau, G., B. Magnini, E. Agirre, P. Vossen and J. Carroll (2002), “MEANING: A Roadmap to Knowledge Technologies”, *Proceedings of COLING Workshop “A Roadmap for Computational Linguistics”*, Taipei, Taiwan.

# Chapter 6

## GERMAN PREPOSITIONS AND THEIR KIN

*A survey with respect to the resolution of PP attachment ambiguities*

Martin Volk

*Stockholm University, Department of Linguistics  
Universitetsvägen 10C, S-10691 Stockholm  
volk@ling.su.se*

**Abstract** This paper surveys German prepositions and their relatives: contracted prepositions, pronominal adverbs, and reciprocal pronouns. We elaborate on corpus frequencies for these and on their properties with respect to PP attachment. Prepositions and contracted prepositions show an overall attachment tendency towards the noun. But pronominal adverbs and reciprocal pronouns show an overall attachment tendency towards the verb and therefore must be treated separately.<sup>1</sup>

**Keywords:** Corpus linguistics, ambiguity resolution, unsupervised learning

### 1. Introduction

Any computer system for natural language processing has to struggle with the problem of ambiguities. If the system is meant to extract precise information from a text, these ambiguities must be resolved. One of the most frequent ambiguities arises from the attachment of prepositional phrases (PPs). A PP that follows a noun (in English or German) can be attached to the noun or to the verb. We did an in-depth study on unsupervised statistical methods to resolve such ambiguities in German sentences based on cooccurrence values derived from a shallow parsed corpus (see (Volk, 2001) and (Volk, 2002)).

Corpus processing consisted of proper name recognition and classification, part-of-speech tagging, lemmatization, phrase chunking, and clause boundary detection. We used a corpus of more than 5 million words from the Computer-Zeitung (CZ), a weekly computer science newspaper. In addition to this train-

ing corpus, we prepared a 3000 sentence corpus with manually annotated syntax trees. From this treebank we extracted over 4000 test cases with ambiguously positioned PPs for the evaluation of the disambiguation method. We will call these test cases the ‘CZ test set’.

As a basis for this study we surveyed German prepositions and their relatives and we checked for prepositions, contracted prepositions, pronominal adverbs and reciprocal pronouns whether they can mutually benefit from each other with respect to attachment tendencies.

## 2. German prepositions

Prepositions in German are a class of words relating linguistic elements to each other with respect to a semantic dimension such as local, temporal, causal or modal. They do not inflect and cannot function by themselves as a sentence unit (cf. (Bußmann, 1990)). But, unlike other function words, a German preposition governs the grammatical case of its argument (genitive, dative or accusative). Frequent German prepositions are *an, für, in, mit, zwischen*.

Prepositions are considered to be a closed word class. Nevertheless it is difficult to determine the exact number of German prepositions. (Schröder, 1990) speaks of “more than 200 prepositions”, but his “Lexikon deutscher Präpositionen” lists only 110 of them. In this preposition dictionary all entries are marked with their case requirement and their semantic features. For instance, *ohne* requires the accusative and is marked with the semantic functions instrumental, modal, conditional and part-of.<sup>2</sup>

The lexical database CELEX (Baayen et al., 1995) contains 108 German prepositions with frequency counts derived from corpora of the “Institut für deutsche Sprache”. This results in the arbitrary inclusion of *nördlich, nordöstlich, südlich* while *östlich* and *westlich* are missing.

Searching through 5.5 million tokens of our tagged computer magazine corpus we found around 540,000 preposition tokens corresponding to 99 preposition types.<sup>3</sup> These counts do not include contracted prepositions. A list of the 75 most frequent German prepositions with frequencies from our corpus can be found in the appendix.

An early frequency count for German (by (Meier, 1964)) lists 18 prepositions among the 100 most frequent word forms. 17 out of these 18 prepositions are also in our top-20 list. Only *gegen* is missing which is on rank 23 in our corpus. This indicates that the usage of the most frequent prepositions is stable over corpora and time.

All frequent prepositions in German have some homograph serving as

- separable verb prefix (e.g. *ab, auf, mit, zu*),
- clause conjunction (e.g. *bis, um*)<sup>4</sup>,

- adverb (e.g. *auf*, *für*, *über*) in often idiomatic expressions (e.g. *auf und davon*, *über und über*),
- infinitive marker (*zu*),
- proper name component (*von*), or
- predicative adjective (e.g. *an*, *auf*, *aus*, *in*, *zu* as in *Die Maschine ist an/aus. Die Tür ist auf/zu.*).

The most frequent homographic functions are separable verb prefix and conjunction. Fortunately, these functions are clearly marked by their position within the clause. A clause conjunction usually occurs at the beginning of a clause, and a separated verb prefix mostly occurs at the end of a clause (*rechte Satzklammer*). A part-of-speech tagger can therefore disambiguate these cases.

Typical (i.e. frequent) prepositions are monomorphemic words (e.g. *an*, *auf*, *für*, *in*, *mit*, *über*, *von*, *zwischen*). Many of the less frequent prepositions are derived or complex. They have turned into prepositions over time and still show traces of their origin. They are derived from other parts-of-speech such as

- nouns (e.g. *angesichts*, *zwecks*),
- adjectives (e.g. *fern*, *unweit*),
- participle forms of verbs (e.g. *entsprechend*, *während*; *ungeachtet*), or
- lexicalized prepositional phrases (e.g. *anhand*, *aufgrund*, *zugunsten*).

German prepositions typically do not allow compounding. It is generally not possible to form a new preposition by a concatenation of prepositions. The two exceptions are *gegenüber* and *mitsamt*. Other concatenated prepositions have led to adverbs like *inzwischen*, *mitunter*, *zwischendurch*.

(Helbig and Buscha, 1998) call the monomorphemic prepositions **primary prepositions** and the derived prepositions **secondary prepositions**. This distinction is based on the fact that only primary prepositions form prepositional objects, pronominal adverbs (cf. section 6.2.2) and prepositional reciprocal pronouns (cf. section 6.2.3).

In addition, this distinction corresponds to different case requirements. The primary prepositions govern accusative (*durch*, *für*, *gegen*, *ohne*, *um*) or dative (*aus*, *bei*, *mit*, *nach*, *von*, *zu*) or both (*an*, *auf*, *hinter*, *in*, *neben*, *über*, *unter*, *vor*, *zwischen*). Most of the secondary prepositions govern genitive (*angesichts*, *beziiglich*, *dank*). Some prepositions (most notably *während*) are in the process of changing from genitive to dative. Some prepositions do not show overt case requirements (*je*, *pro*, *per*; cf. (Schaeder, 1998)) and are used with determiner-less noun phrases.

Some prepositions show other idiosyncracies. The preposition *bis* often takes another preposition (*in*, *um*, *zu* as in 6.2) or combines with the particle *hin* plus a preposition (as in 6.2). The preposition *zwischen* is special in that it requires a plural argument (as in 6.2), often realized as a coordination of NPs (as in 6.2).

- . Portables mit 486er-Prozessor werden ***bis zu 20 Prozent*** billiger.
- . ... und berücksichtigt auch Daten und Datentypen ***bis hin zu Arrays*** oder den Records im VAX-Fortran.
- . Die Verbindungstopologie ***zwischen den Prozessoren*** lässt sich als dreidimensionaler Torus darstellen.
- . Durch Microsoft Access müssen sich die Anwender nicht mehr länger ***zwischen Bedienerfreundlichkeit und Leistung*** entscheiden.

**Results for PP attachment** We explored various possibilities to extract PP disambiguation information from the automatically annotated CZ corpus. We first used it to gather frequency data on the cooccurrence of pairs: nouns + prepositions and verbs + prepositions.

The cooccurrence value is the ratio of the bigram frequency count  $freq(word, preposition)$  divided by the unigram frequency  $freq(word)$ . For our purposes *word* can be the verb *V* or the reference noun  $N_1$ . The ratio describes the percentage of the cooccurrence of *word + preposition* against all occurrences of *word*. It is thus a straightforward association measure for a word pair. The cooccurrence value can be seen as the attachment probability of the preposition based on maximum likelihood estimates. We write:

$$cooc(W, P) = freq(W, P) / freq(W)$$

with  $W \in \{V, N_1\}$ . The cooccurrence values for verb *V* and noun  $N_1$  correspond to the probability estimates by (Ratnaparkhi, 1998) except that Ratnaparkhi includes a back-off to the uniform distribution for the zero denominator case. We added special precautions for this case in our disambiguation algorithm. The cooccurrence values are also very similar to the probability estimates by (Hindle and Rooth, 1993).

We started by computing the cooccurrence values over word forms for nouns, prepositions, and verbs based on their part-of-speech tags. In order to compute the pair frequencies  $freq(N_1, P)$ , we search the training corpus for all token pairs in which a noun is immediately followed by a preposition. The treatment of verb + preposition cooccurrences is different from the treatment of N+P pairs since verb and preposition are seldom adjacent to each other

in a German sentence. On the contrary, they can be far apart from each other, the only restriction being that they cooccur within the same clause. We use the clause boundary information in our training corpus to enforce this restriction. For computing the cooccurrence values we accept only verbs and nouns with an occurrence frequency of more than 10.

With the N+P and V+P cooccurrence values for word forms we did a first evaluation over the CZ test set with the following simple disambiguation algorithm.

```
if ( cooc(N1,P) && cooc(V,P) ) then
  if ( cooc(N1,P) >= cooc(V,P) ) then
    noun attachment
  else
    verb attachment
```

We found that we can only decide 57% of the test cases with an accuracy of 71.4% (93.9% correct noun attachments and 55.0% correct verb attachments). This shows a striking imbalance between the noun attachment accuracy and the verb attachment accuracy. This imbalance was countered with a noun factor which was automatically derived from the corpus based on the overall attachment tendency of prepositions towards nouns in comparison to their tendency towards verbs (cf. (Volk, 2002)). This move leads to an improvement of the overall attachment accuracy to 81.3%. We then went on to lemmatize all word forms which also included mapping contracted prepositions to their corresponding bare forms.

## 2.1 Contracted Prepositions

Certain German primary prepositions combine with a determiner to contracted forms. This process is restricted to the prepositions *an*, *auf*, *ausser*, *bei*, *durch*, *für*, *hinter*, *in*, *neben*, *über*, *um*, *unter*, *von*, *vor*, *zu*. Our corpus contains about 89,000 tokens that are tagged as contracted prepositions (14% of all preposition tokens). The contracted form stands usually for a combination of the preposition with the definite determiner *der*, *das*, *dem*.<sup>5</sup> If a contracted preposition is available, it will not always substitute the separate usage of preposition and determiner but rather compete with it. For example, the contracted preposition *beim* (example 6.2.1) is used as separate forms with a definite determiner in 6.2.1. Example 6.2.1 shows a sentence with *bei* plus an indefinite determiner. But the usage of the contracted preposition would also be possible (*Beim Ausfall einer gesamten CPU*), and we claim that it would not change the meaning. This indicates that sometimes the contracted preposition might stand for a combination of the preposition with the indefinite determiner *einer*, *ein*, *einem*.

- . Detlef Knott, Vertriebsleiter ***beim Softwarehaus Computenz GmbH*** ...
- . Eine adäquate Lösung fand sich ***bei dem indischen Softwarehaus CMC***, das ein Mach Plan-System bereits ... in die Praxis umgesetzt hatte:
- . ***Bei einem Ausfall*** einer gesamten CPU springt der Backup-Rechner für das ausgefallene System in die Bresche.

For the most frequent contracted prepositions (*im, zum, zur, vom, am, beim, ins*), the separate usage of determiner and preposition indicates a special stress on the determiner. The definite determiner then resembles a demonstrative pronoun.

The less frequent contracted prepositions sound colloquial (e.g. *aufs, überm*). The frequency overview in the appendix shows that these contracted prepositions are more often used in separated than in contracted form in our newspaper corpus. (Helbig and Buscha, 1998) (p. 388) claim that *ans* is unmarked (“völlig normalsprachlich”), but our frequency counts contradict this claim. In our newspaper corpus *ans* is used 199 times but *an das* occurs 611 times. This makes *ans* the borderline case between the clearly unmarked contracted prepositions and the ones that are clearly marked as colloquial in written German.

Some contracted prepositions are required by specific constructions in standard German and should be treated separately with respect to PP attachment. Among these are (according to (Drosdowski, 1995)):

- *am* with the superlative: *Sie tanzt am besten.*
- *am* or *beim* with infinitives that are used as nouns: *Er ist am Arbeiten. Er ist beim Kochen.*
- *am* as a fixed part of date specifications: *Er kommt am 15. Mai.*

By using verb lemmas and noun lemmas (including noun decompounding; i.e. using only the last component of a compound noun), and by mapping contracted prepositions to their bare preposition counterparts, we increased the coverage of our disambiguation procedure from 57% to 83% of the test cases with only a minor loss in accuracy which could not be attributed to the contracted prepositions but rather to low frequencies and idiosyncracies of some verbs and nouns. It is therefore safe to conclude that contracted prepositions can be dealt with in the same way as base prepositions for the PP attachment task.

The base prepositions in our test set (3831 tokens) display a tendency towards noun attachment (63%) rather than verb attachment (37%). The contracted prepositions in the test set (640 tokens) display a similar, slightly weaker tendency towards noun attachment (55%).

## 2.2 Pronominal Adverbs

In another morphological process primary prepositions can be embedded into pronominal adverbs. A pronominal adverb is a combination of a particle (*da(r)*, *hier*, *wo(r)*) and a preposition resulting in (e.g. *daran*, *dafür*, *hierunter*, *woran*, *wofür*).<sup>6</sup> In colloquial German pronominal adverbs with *dar* are often reduced to *dr*-forms (e.g. *dran*, *drin*, *drunter*), and we found some dozen occurrences of these in our corpus.

Pronominal adverbs are used to substitute and refer to a prepositional phrase. The forms with *da(r)* are often used in place holder constructions, where they serve as (mostly cataphoric) pointers to various types of clauses.

- . **Cataphoric pointer to a *daf*-clause:** Es sollte *darauf* geachtet werden, daß auch die Hersteller selbst vergleichbar sind.
- . **Cataphoric pointer to an infinitive clause:** Die Praxis der Software-Nutzungsverträge zielt *darauf* ab, den mitunter gravierenden Wandel in den DV-Strukturen eines Unternehmens nicht zu behindern ...
- . **Anaphoric pointer to a noun phrase:** Vielmehr können sich /36-Kunden, die den Umstieg erst später wagen wollen, mit der RPG II 1/2 *darauf* vorbereiten.

The complete list of pronominal adverbs can be found in the appendix. It is striking that the frequency order of this list does not correspond to the frequency order of the preposition list. The most frequent prepositions *in* and *von* are represented only on ranks 13 and 6 in the pronominal adverb list. Obviously, pronominal adverbs behave differently from their corresponding prepositions. Pronominal adverbs can only substitute prepositional complements (as in 6.2.2) with the additional restriction that the PP noun must not be an animate object (as in 6.2.2; the asterisk marking the ungrammatical variant). Pronominal adverbs cannot substitute adjuncts. Those will be substituted by adverbs that represent their local (*hier*, *dort*; see 6.2.2) or temporal character (*damals*, *dann*). (de Lima, 1997) exploits these facts to automatically determine verbal subcategorisation frames based on unambiguously positioned pronominal adverbs in main clauses.

- . Die Wasserchemiker warten *auf solche Geräte / darauf* ...
- . Absolut neue Herausforderungen warten *auf die Informatiker / \*darauf / auf sie* beim Stichwort “genetische Algorithmen” ...
- . Daher wird *auf dem Börsenparkett / \*darauf / dort* heftig über eine mögliche Übernahme spekuliert.

We restrict pronominal adverbs to combinations of the above-mentioned particles (*da*, *hier*, *wo*) with prepositions. Sometimes other combinations with prepositions are included as well. The guidelines for the German tag set STTS (Schiller et al., 1995) includes combinations with *des* and *dem* (*deswegen*;

*ausserdem, trotzdem*; also with postpositions: *demgemäß, demzufolge, demgegenüber*).

On the other hand the STTS separates the combinations with *wo* into the class of adverbial interrogative pronouns. This classification is appropriate for the purpose of part-of-speech tagging. The distributional properties of *wo*-combinations are more similar to other interrogative pronouns like *wann* than to regular pronominal adverbs. But for the purpose of investigating prepositional attachments, we will concentrate on those pronominal adverbs that behave most similar to PPs.

Our test set contains 152 test cases with pronominal adverbs. 81% of these cases are verb attachments. This contrasts sharply with the 60% noun attachments that we observed over all the prepositional test cases. This makes it obvious that pronominal adverbs require special treatment with respect to their attachment and cannot be resolved by using cooccurrence values derived from prepositions. By computing cooccurrence values over the pronominal adverbs in our corpus we were able to improve the attachment accuracy for pronominal adverbs to about 85%.

## 2.3 Reciprocal Pronouns

Yet another disguise of primary prepositions is their combination with the reciprocal pronoun *einander*.<sup>7</sup> The preposition and the pronoun constitute an orthographic unit which substitutes a prepositional phrase.

A reciprocal pronoun may modify a noun (as in example 6.2.3) or a verb (as in 6.2.3). Most reciprocal pronouns can also be used as nouns (see 6.2.3); some are nominalized so often that they can be regarded as lexicalized (e.g. *Durcheinander, Miteinander, Nebeneinander*).

- . ... und damit eine Modellierung von Objekten der realen (Programmier-) Welt und ihrer Beziehungen **untereinander** darstellen können.
- . Ansonsten dürfen die Behörden nur die vom Verkäufer und vom Erwerber eingegangenen Informationen **miteinander** vergleichen.
- . Chaos ist in der derzeitigen Panik- und Krisenstimmung nicht nur ein Wort für wildes **Durcheinander**, sondern ...

In our corpus we found 16 different reciprocal pronouns with prepositions. The frequency ranking is listed in the appendix. It is striking that some of the P+*einander* combinations are more frequent than the reciprocal pronoun itself.

With respect to their attachment reciprocal pronouns are similar to pronominal adverbs in that they show a strong tendency towards verb attachment. We checked through our treebanks and found 34 reciprocal pronouns. Four of these were noun attachments (12%) including one deverbal noun (*Umgehen miteinander*), and two were adjective attachments again including one dever-

bal adjective (present participle; *nebeneinander liegenden*). This leaves 28 cases (82%) for verb attachment.

## 2.4 Prepositions in Other Morphological Processes

Some prepositions are subject to conversion processes. Their homographic forms belong to other word classes. In particular, there are  $P_1 + \text{conjunction} + P_2$  sequences (*ab und zu*, *nach wie vor*, *über und über*) that are idiomized and function as adverbials (cf. example 6.2.4). They are derived from prepositions but they do not form PPs. As long as they are symmetrical, they can easily be recognized. All others are best listed in a lexicon so that they are not confused with coordinated prepositions.

Some such coordinated sequences must be treated as  $N + \text{conjunction} + N$  (*das Auf und Ab*, *das Für und Wider*; cf. 6.2.4) and are also outside the scope of our research. Finally, there are few prepositions that allow a direct conversion to a noun such as *Gegenüber* in 6.2.4.

- . Eine Vielzahl von Straßennamensänderungen wird ***nach und nach*** noch erfolgen.
- . Nachdem sie das ***Für und Wider*** gehört haben, können die Zuschauer ihre Meinung ... kundtun.
- . Verhandlungen enden häufig in der Sackgasse, weil kein Verhandlungspartner sich zuvor Gedanken über die Situation seines ***Gegenübers*** gemacht hat.

Prepositions are often used to form adverbs. We have already mentioned that P+P compounds often result in adverbs (e.g. *durchaus*, *nebenan*, *überaus*, *vorbei*). Even more productive is the combination with the particles *hin* and *her*. They are used as suffix *nachher*, *vorher*, *mithin*, *ohnehin* or as prefix *herauf*, *herüber*, *hinauf*, *hiniüber*. These adverbs are sometimes called prepositional adverbs (cf. (Fleischer and Barz, 1995)). The particles can also combine with pronominal adverbs (*daraufhin*).

In addition, there is a limited number of preposition combinations with nouns (*bergauf*, *kopfüber*, *tagsüber*) and adjectives (*hellauf*, *rundum*, *weitaus*) that function as adverbs if the preposition is the last element. Sometimes the preposition is the first element, which leads to a derivation within the same word class (*Ausfahrt*, *Nachteil*, *Vorteil*, *Nebensache*).

Finally, most of the verbal prefixes can be seen as preposition + verb combinations. Some of them function only as separable prefix (*ab*, *an*, *auf*, *aus*, *bei*, *nach*, *vor*, *zu*), others can be separable or inseparable (*durch*, *über*, *um*, *unter*). Note that the meaning contribution of the preposition to the verb varies as much as the semantic functions of the preposition. Consider for example the preposition *über* in *überblicken* (to survey; literally: to view over), *übersehen* (to overlook, to disregard, to realize; literally: to look over or to look away), and *übertreffen* (to surpass; literally: to aim better).

The preposition *mit* shows an idiosyncratic behaviour when it occurs with prefixed verbs (be they separable as in 6.2.4 or inseparable as in 6.2.4). In this case *mit* does not combine with the verb but rather functions as an adverb.<sup>8</sup>

- . Schröder ist seit 22 Jahren für die GSI-Gruppe tätig und hat die deutsche Dependance **mit** aufgebaut.
- . Die Hardwarebasis soll noch erweitert werden und andere Unix-Plattformen **mit** einbeziehen.

This analysis is shared by (Zifonun et al., 1997) (p. 2146). *mit* can function like a PP-specifying adverb (see 6.2.4). And in example 6.2.4 it looks more like a stranded separated prefix (cf. *an Bord mitzunehmen*). (Zifonun et al., 1997) note that the distribution of *mit* differs from full adverbs. It is rather similar to the adverbial particles *hin* and *her*. All of them can only be moved to the *Vorfeld* in combination with the constituent that they modify.

- . ... und deren Werte **mit** in die DIN 57848 für Bildschirme eingingen.
- . ... geht man dazu über, Subunternehmer **mit** an Bord zu nehmen.

## 2.5 Postpositions and Circumpositions

In terms of language typology German is regarded as a preposition language while others, like Japanese or Turkish, are postposition languages. But in German there are also rare cases of postpositions and circumpositions. Circumpositions are discontinuous elements consisting of a preposition and a “postpositional element”. This postpositional element can be an adverb (as in example 6.2.5) or a “preposition form” (as in example 6.2.5). Even pronominal adverbs can take postpositional elements to form circumpositional phrases (see example 6.2.5).

- . Beispielsweise können Werte und Grafiken in ein Textdokument exportiert oder Messungen **aus einer Datenbank heraus** parametriert und gestartet werden.
- . ... oder **vom Programm aus** direkt gestartet werden.
- . Die Messegesellschaft hat **dariüber hinaus** globale Netztechnologien und verschiedene Endgeräte in dieser Halle angesiedelt.

The case of postpositions is similar. There are few true postpositions (e.g. *halber*, *zufolge*; see 6.2.5), but others are homographic with prepositions (see *nach*, *über*, which are mostly used as prepositions, functioning as postpositions in the examples 6.2.5 and 6.2.5).

- . Über die Systems in München werden **Softbank-Insidern zufolge** Gespräche geführt.
- . Das größte Potential für die Branche steckt **seiner Ansicht nach** in der Verknüpfung von Firmen.

- . Und das bleibt auch **die Woche über** so.

Because of these homographs the correct part-of-speech tagging for postpositions and postpositional elements of circumpositions is a major problem. It works correctly if the subsequent context is prohibitive for the preposition reading (e.g. when the postposition is followed by a verb). But in other pre vs. post ambiguities a part-of-speech tagger often fails since the preposition reading is so dominant for these words. Special correction rules are needed.

### 3. Conclusions

This survey has focused on German prepositions and their relatives: contracted prepositions, pronominal adverbs, reciprocal pronouns, circumpositions and postpositions. Based on automatically annotated corpora and a small treebank we have investigated their behaviour with respect to noun versus verb attachment. We have found that

- contracted prepositions can be mapped to prepositions since they display similar attachment tendencies (towards noun attachment),
- prepositions behave different from pronominal adverbs and reciprocal pronouns (tendency towards verb attachment), and
- a number of prepositional idiosyncrasies can be exploited for *am*, *bis*, *mit*, *zwischen* and others.

Since we did a quantitative evaluation, we only evaluated 59 preposition types because our test set happened to contain only these prepositions. A thorough evaluation of the attachment tendencies of the remaining 40 plus prepositions needs to be tackled next (together with those prepositions that occurred only rarely in the test set and the few missing contracted forms and pronominal adverbs).

### Notes

1. This paper is based on my research at the University of Zurich in a project supported by the Swiss National Science Foundation under grant 12-54106.98.
2. See also (Klaus, 1999) for a detailed comparison of the range of German prepositions as listed in a number of recent grammar books.
3. These figures are based on automatically assigned part-of-speech tags. If the tagger systematically mistagged a preposition, the counting procedure does not find it. In the course of the project we realized that this happened to the prepositions *a*, *via* and *voller* as used in the following example sentences (all examples in this paper are from the Computer-Zeitung, Konradin-Verlag, 1993-1997).
  - . Derselbe Service in der Regionalzone (bis zu 50 Kilometern) kostet 23 Pfennig **a 60 Sekunden**.
  - . Master und Host kommunizieren **via IPX**.
  - . Windows steckt **voller eigener Fehler**.
4. (Jaworska, 1999) (p. 306) argues that “clause-introducing preposition-like elements are indeed prepositions”.

5. (Helbig and Buscha, 1998) (p. 388) mention that it is possible to build contracted forms with the determiner *den*: *hintern*, *übern*, *untern*. But these forms are very colloquial and do not occur in our corpus.
6. This is why pronominal adverbs are sometimes called prepositional adverbs (e.g. in (Zifonun et al., 1997)) or even prepositional pronouns (e.g. in (Langer, 1999)).
7. Sometimes the word *gegenseitig* is also considered to be a reciprocal pronoun. Since the preposition *gegen* in this form cannot be substituted by any other preposition, we take this to be a special form and do not discuss it here.
8. A detailed study of the preposition *mit* can be found in (Springer, 1987).

## References

Baayen, R. H., Piepenbrock, R., and van Rijn, H. (1995). The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania.

Bußmann, H. (1990). *Lexikon der Sprachwissenschaft*. Kröner Verlag, Stuttgart, 2. revised edition.

de Lima, E. F. (1997). Acquiring German prepositional subcategorization frames from corpora. In Zhou, J. and Church, K., editors, *Proc. of the Fifth Workshop on Very Large Corpora*, pages 153–167, Beijing and Hongkong.

Drosdowski, G., editor (1995). *DUDEN. Grammatik der deutschen Gegenwartssprache*. Bibliographisches Institut, Mannheim, 5. edition.

Fleischer, W. and Barz, I. (1995). *Wortbildung der deutschen Gegenwartssprache*. Niemeyer, Tübingen, 2. edition.

Helbig, G. and Buscha, J. (1998). *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Langenscheidt. Verlag Enzyklopädie, Leipzig, Berlin, 18. edition.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Jaworska, E. (1999). Prepositions and prepositional phrases. In Brown, K. and Miller, J., editors, *Concise Encyclopedia of Grammatical Categories*, pages 304–311. Elsevier, Amsterdam.

Klaus, C. (1999). *Grammatik der Präpositionen: Studien zur Grammatikographie; mit einer thematischen Bibliographie*, volume 2 of *Linguistik International*. Peter Lang, Frankfurt.

Langer, H. (1999). *Parsing-Experimente*. Habilitationsschrift, Universität Osnabrück.

Meier, H. (1964). *Deutsche Sprachstatistik*. Georg Olms Verlag, Hildesheim.

Ratnaparkhi, A. (1998). Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of COLING-ACL-98*, Montreal.

Schaeder, B. (1998). Die Präpositionen in Langenscheidts Großwörterbuch Deutsch als Fremdsprache. In Wiegand, H. E., editor, *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von "Langenscheidts Großwörterbuch Deutsch als Fremdsprache"*, volume 86

of *Lexicographica. Series Maior*, pages 208–232. Niemeyer Verlag, Tübingen.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textcorpora mit STTS (Draft). Technical report, Universität Stuttgart. Institut für maschinelle Sprachverarbeitung.

Schröder, J. (1990). *Lexikon deutscher Präpositionen*. Verlag Enzyklopädie, Leipzig.

Springer, D. (1987). *Valenz der Verben mit präpositionalem Objekt “von”, “mit” : eine kontrastive Studie*. Wydawnictwo Wyższej Szkoly Pedagogicznej, Zielona Gora.

Volk, M. (2001). *The automatic resolution of prepositional phrase attachment ambiguities in German*. Habilitationsschrift, University of Zurich.

Volk, M. (2002). Combining unsupervised and supervised methods for PP attachment disambiguation. In *Proc. of COLING-2002*, Taipeh.

Zifonun, G., Hoffmann, L., and Strecker, B. (1997). *Grammatik der deutschen Sprache*, volume 7 of *Schriften des Instituts für deutsche Sprache*. de Gruyter, Berlin.

## Appendix: Prepositions

This appendix lists the 75 most frequent prepositions of the Computer-Zeitung (1993-95+1997). We have added the classification as either primary or secondary preposition. Furthermore we have added the case requirement (accusative, dative, genitive), contracted forms that occur in our corpus, and pronominal adverb forms. Pure postpositions and circumpositions are not listed.

rank	preposition	frequency	type	case	contr.	pron. adv.
1	in	84662	prim.	acc/dat	im/ins	darin
2	von	71685	prim.	dat	vom	davon
3	für	64413	prim.	acc	fürs	dafür
4	mit	61352	prim.	dat		damit
5	auf	49752	prim.	acc/dat	aufs	darauf
6	bei	27218	prim.	dat	beim	dabei
7	über	19182	prim.	acc/dat	überm/s	darüber
8	an	18256	prim.	acc/dat	am/ans	daran
9	zu	17672	prim.	dat	zum/zur	dazu
10	nach	15298	prim.	dat		danach
11	aus	13949	prim.	dat		daraus
12	durch	12038	prim.	acc	durchs	dadurch
13	bis	11253	sec.	acc		
14	unter	10129	prim.	acc/dat	unterm/s	darunter
15	um	9880	prim.	acc	ums	darum
16	vor	9852	prim.	acc/dat	vorm/s	davor
17	zwischen	5079	prim.	acc/dat		dazwischen
18	seit	4194	sec.	dat		(seitdem)
19	pro	4175	sec.	/		
20	ohne	3007	prim.	acc		
21	neben	2733	prim.	acc/dat		daneben
22	laut	2438	sec.	dat		
23	gegen	2127	prim.	acc		dagegen
24	per	2011	sec.	/		
25	ab	1884	sec.	acc/dat		
26	gegenüber	1707	sec.	dat		
27	innerhalb	1509	sec.	gen		
28	trotz	1260	sec.	dat/gen		(trotzdem)
29	wegen	1048	prim.	dat/gen		(deswegen)
30	aufgrund	949	sec.	gen		
31	während	747	sec.	dat/gen		(w.-dessen)
32	hinter	721	prim.	acc/dat	hinterm/s	dahinter
33	statt	611	sec.	gen		(s.-dessen)
34	angesichts	553	sec.	gen		
35	außer	446	sec.	dat		(außerdem)
36	dank	414	sec.	dat/gen		
37	je	390	sec.	/		
38	mittels	380	sec.	dat/gen		
39	hinsichtlich	354	sec.	gen		
40	namens	341	sec.	gen		

rank	preposition	frequency	type	case	contr.	pron. adv.
41	außerhalb	310	sec.	gen		
42	inklusive	293	sec.	gen		
43	einschließlich	284	sec.	gen		
44	anhand	258	sec.	gen		
45	samt	164	sec.	dat		
46	gemäß	153	sec.	dat/gen		
47	bezüglich	148	sec.	gen		
48	zugunsten	136	sec.	gen		
49	anlässlich	132	sec.	gen		
50	binnen	120	sec.	dat/gen		
51	anstelle	105	sec.	gen		
52	infolge	103	sec.	gen		(i.-dessen)
53	seitens	95	sec.	gen		
54	jenseits	90	sec.	gen		
55	entgegen	76	sec.	dat		
56	entlang	64	sec.	acc/gen		
57	unterhalb	58	sec.	gen		
58	anstatt	56	sec.	gen		
59	nahe	49	sec.	gen		
60	mangels	44	sec.	gen		
61	seiten	39	sec.	gen		
62	versus	32	sec.	gen		
63	nebst	31	sec.	dat		
64	wider	26	sec.	acc		
65	oberhalb	23	sec.	gen		
66	ob	21	sec.	gen		darob
67	mitsamt	21	sec.	dat		
68	ungeachtet	20	sec.	gen		
69	abseits	20	sec.	gen		
70	zuzüglich	18	sec.	gen		
71	zwecks	17	sec.	gen		
72	ähnlich	15	sec.	gen		
73	inmitten	12	sec.	gen		
74	eingangs	9	sec.	gen		
75	südlich	8	sec.	gen		
...	...					

## Appendix: Contracted Prepositions

This appendix lists all contracted prepositions of the Computer-Zeitung (1993-95+1997). The table includes contracted forms for the prepositions *an, auf, bei, durch, für, hinter, in, über, um, unter, von, vor, zu*. In order to illustrate the usage tendency we added the frequencies for the non-contracted forms.

rank	contr. prep.	freq.	prep. + det.	freq.	prep. + det.	freq.
1	im	40940	in dem	857	in einem	2365
2	zum	14225	zu dem	330	zu einem	1578
3	zur	13537	zu der	219	zu einer	986
4	vom	6299	von dem	534	von einem	1061
5	am	6136	an dem	442	an einem	506
6	beim	4641	bei dem	551	bei einem	759
7	ins	2155	in das	1053	in ein	521
8	ans	199	an das	611	an ein	171
9	fürs	154	für das	3787	für ein	879
10	aufs	125	auf das	1281	auf ein	600
11	übers	109	über das	1598	über ein	684
12	ums	60	um das	302	um ein	372
13	durchs	53	durch das	645	durch ein	373
14	unterm	36	unter dem	1062	unter einem	102
15	unders	10	unter das	27	unter ein	6
16	vors	4	vor das	20	vor ein	44
17	hinterm	4	hinter dem	102	hinter einem	5
18	überm	2	über dem	142	über einem	50
19	vorm	1	vor dem	598	vor einem	263
20	hinters	1	hinter das	3	hinter ein	0

## Appendix: Pronominal Adverbs

This appendix lists all pronominal adverbs of the Computer-Zeitung (1993-95+1997) sorted by the cumulated frequency of the corresponding preposition.

rank	prep.	freq.	da-form	freq.	hier-form	freq.	wo-form	freq.
1	bei	6929	dabei	5861	hierbei	381	wobei	687
2	mit	6446	damit	6332	hiermit	36	womit	78
3	zu	3508	dazu	3099	hierzu	348	wozu	61
4	für	2767	dafür	2410	hierfür	309	wofür	48
5	von	1777	davon	1708	hiervon	20	wovon	49
6	über	1783	darüber	1766	hierüber	5	worüber	12
7	durch	1601	dadurch	1385	hierdurch	54	wodurch	162
8	gegen	1420	dagegen	1397	hiergegen		wogegen	23
9	auf	1324	darauf	1267	hierauf	19	worauf	38
10	an	789	daran	737	hieran	9	woran	43

rank	prep.	freq.	da-form	freq.	hier-form	freq.	wo-form	freq.
11	in	738	darin	685	hierin	18	worin	35
12	nach	613	danach	531	hiernach	3	wonach	79
13	unter	601	darunter	587	hierunter	6	worunter	8
14	aus	463	daraus	432	hieraus	18	woraus	13
15	um	377	darum	367	hierum		worum	10
16	neben	331	daneben	331	hierneben		woneben	
17	vor	148	davor	146	hievor		wovor	2
18	hinter	135	dahinter	135	hierhinter		wohinter	
19	zwischen	26	dazw.	26	hierz.		wozw.	

All primary prepositions are represented except for *ohne* and *wegen*. Queries to the internet search engine Google reveal that pronominal adverb forms for *wegen* do exist albeit with low frequencies (*dawegen* 8, *hierwegen* 82, *wowegen* 3!). The internet search engine also finds examples for those forms with zero frequency in the Computer-Zeitung (*hiergegen* being by far the most frequent form).

## Appendix: Reciprocal Pronouns

This appendix lists all prepositional reciprocal pronouns of the Computer-Zeitung (1993-95+1997). The table includes the pure pronoun *einander* (rank 7).

rank	reciprocal pronoun	frequency
1	miteinander	609
2	untereinander	187
3	voneinander	161
4	aufeinander	91
5	auseinander	66
6	nebeneinander	58
7	<i>einander</i>	47
8	zueinander	43
9	gegeneinander	37
10	hintereinander	28
11	nacheinander	20
12	durcheinander	14
13	aneinander	13
14	ineinander	12
15	beieinander	12
16	übereinander	7
17	füreinander	1

Five primary prepositions do not have reciprocal pronouns in this corpus. But for all of them we find usage examples in the internet (with *wegeneinander* being the least frequent).

# Chapter 7

## DIRECTIONALITY SELECTION

Marcus Kracht \*

*Department of Linguistics, UCLA, PO Box 951543, 405 Hilgard Avenue, Los Angeles, CA 90095–1543, USA*

[kracht@humnet.ucla.edu](mailto:kracht@humnet.ucla.edu)

**Abstract** It has frequently been observed that locative PPs are bimorphemic, consisting of two heads: one specifying the location, and the other one specifying the path or directionality. This bipartite structure carries over to other PPs (predicatives and habitives). This structure is problematic for standard syntactic theory. For ordinary selection of those locatives consists in the selection of *two* heads rather than one, contrary to theory. On the other hand, there is a kind of selection that selects just one of them, namely the outer one, which specifies directionality. This is the **directionality selection** that is the topic of this paper. We shall study this type of selection in various languages. It will emerge that directionality selection is not at all marginal, and that it is responsible for systematic differences between various languages.

**Keywords:** locatives, selection, syntax, directionality

### 1. Introduction

The proper understanding of the way space is encoded in language is of extreme importance. Moreover, language is filled with expressions that originate one way or another in spatial talk. Whenever a language has a rich case system it is because it has plenty of local cases. Languages which have few cases, on the other hand, do have adpositions that fulfil the same function (English, French and German are a case in point). It turns out that the mechanics of the PPs is the same as that of the local cases.

In the last years, space and spatial expressions have received growing interest (see for example the collection (Bloom et al 1989), (Jackendoff, 1983),

\*I wish to thank Raimo Anttila and two anonymous referees for helping me improve this paper.

(Wunderlich et al. 1986), (Svorou, 1993), (Maienborn, 2001), (Fong 1997), (Levinson, 2003) and (Kracht 2002)). Primarily, the emphasis has been on the study of locations and location denoting expressions or on the metaphorical use of space. We have stressed in (Kracht 2003) that the mechanics of directionality is an integral part of locatives, something which is often neglected in favour of the purely locational aspect.

In this paper I shall be concerned with locative expressions and the interaction of syntax and semantics. Locatives have the following structure (order irrelevant):

(7.1)	[from	[behind	[the car]]]
	[M	[L	DP]]

We call DP the **landmark**, L the **localiser**, and M the **modalizer**. L+DP is the **location phrase**, M+L+DP the **mode phrase**. Semantically, the landmark contributes an object (car), which may or may not move in space. L+DP returns a set of spatial regions ('neighbourhoods'), which may change through time (behind the car). Finally, M+L+DP describes the way in which a certain element changes its position with respect to this (possibly changing) neighbourhood (from behind the car).

Directionality was studied from a semantic point of view by (fong:locatives). Fong argues that directionals denote phase quantifiers in the sense of (loebner:wahrnfalsch), and that verbs may either denote a single phase (in which case they are static) or two successive phases. Directionals either specify a property of the first phase (coinitial), or of the second phase (cofinal). Fong views the phases as completely formal objects, which allows verbs to select directionals even when no change in state or location occurs. This approach turns the exact directional meaning of the directionals into a mystery (see (kracht:locatives)). Instead, we have proposed that the directional meaning is removed upon selection. (An inverse scenario, that directional meaning is added upon selection, is also conceivable, but I see no way to implement it.) Yet, this argument, although workable, ignores that the particular choice of directionals in Finnish is to a large extent predictable.

## 2. Modes

The meanings of modalisers are called **modes**. In the literature there is no consensus on the name for these meanings; typically, modes express properties of the motion of the trajector. So, they can often — but not always — be viewed as modifying the **path** of the trajector. There are several basic modes (see (Melcuk, 1994): **static** (the object is at rest inside the neighbourhood during event time), **cofinal** (the object moves into the location), **coinitial** (the object moves out of the location), **transitory** (the object moves into and then out of the location), **approximative** (the object approaches the location), **recessive**

(the object goes away from the location). The object of which the directional asserts change of location is called the **mover** (some call it **trajector**). We have argued in (kracht:locatives) that static locatives indicate event location rather than participant location.

Our paper is mainly based on Finno–Ugric languages, with comparison to some Indo–European languages. Proto–Finno–Ugric is said to have distinguished by means of cases only mode, not location. It had three grammaticalized modes: static, cofinal and coinitial. This threefold distinction is clearly visible still today. It should be noted that (as in many other languages), mode heads are not only used to derive spatial (locative) expressions; they naturally expand into other cognitive domains; for example, they can typically also denote predicative expressions, and habitives. (Alhoniemi, 1967), based on work by Paavo Siro, has studied the meaning of locatives in Finnish and Cheremiss (today called Mari). He gives the following table of cases, where items in the same row have identical mode:

Locative cases	Predicative cases	Habitive–locative cases
Inessive	Essive	Adessive
Illative	Translative	Allative
Elative	Elative	Ablative

In Finnish grammar the cases of the first column are called **inner locatives**, the ones in the last **outer locatives**. Notice that the outer locatives serve a dual purpose: on the one hand they are locatives (*talolla* ‘at the house’) on the other hand they denote possession (*minulla* on *talo* ‘I have a house’). For example, the Finnish essive *laivana* means ‘being a ship’ or ‘as a ship’, the transformative *laivaksi* means ‘transforming (changing) into a ship’. Notice that the third entry in this column is the elative, originally a locative case, but used in many other connections, too, for example, as a substitute for the partitive.

It is irrelevant for syntactic and semantic purposes in which way these elements are realized (that is whether they are cases, nouns, or adpositions). We have shown that within one language, local DPs and local PPs are syntactically and semantically alike, only their morphology is different. For example, English has no cases, and the locatives are mainly realized through prepositions. However, there are subtle details. First, the distinction between static and cofinal has become marginalised. On the other hand, it still exists in the pair *in/into*(and *on/onto*). (In colloquial speech, this distinction is less and less observed.) The coinitial counterpart is *out*, which selects the genitive (realized by *of*). In German the contrast static/cofinal is encoded by the dative versus accusative on the DP (*an der Wand* ‘on the wall’, *an die Wand* ‘onto the wall’). Finnish and Hungarian both have a fair amount of local cases. For example, Finnish has six cases, corresponding — roughly — to the trias *in/into/out of* and *at/to/from*. Hungarian adds *on/onto/from onto*. In

these languages, other Ls (= localisers) are expressed by means of adpositions (for example Hu. *alatt* ‘under’), and it is possible to coordinate a locative DP with a locative PP. We shall therefore consider the realization immaterial. This is why we shall talk also of an allative in German (realized by *an+ACC*). Cases are such markers that are selected by a higher head. Notice that in German (and for example in many Indo-Aryan languages) as many as three (or even four) elements make up the marking of a DP (see (masica:indoaryan)). These are (a) M, (b) L, and (c) the case of the DP (which in Hindi is once again a postposition governing oblique case).

### 3. One Word — Three Meanings

We shall outline the basic analysis from (Kracht, 2003). Language is a set of signs, and a **grammar** is a language together with a family of operations. A **sign** is a triple  $\langle e, c, m \rangle$ ,  $e$  being the **exponent** (typically a string),  $c$  the **category** (formed from attribute value structures (AVSs) using directional slashes) and  $m$  its **meaning** (typically a typed  $\lambda$ -term). For example,

$$(7.2) \quad \text{MAN} = \langle \text{man}, N, \lambda x. \text{man}'(x) \rangle$$

s a sign of English (simplifying matters greatly). Another sign is

$$(7.3) \quad A = \langle a, DP/N, \lambda P. \lambda Q. \exists x. P(x) \wedge Q(x) \rangle$$

There is a binary operation ‘ $\circ$ ’ (called **merge**) which on the side of exponents concatenates the strings (with a blank interspersed), on the side of categories applies slash–cancellation (according to the rules  $\alpha/\beta \cdot \beta = \alpha$ , and  $\beta \cdot \alpha/\beta = \alpha$ ) and on the side of meanings applies the functor to its argument. Thus,  $A \circ \text{MAN}$  is a sign and we have

$$(7.4) \quad A \circ \text{MAN} = \langle a \text{ man}, DP, \lambda Q. \exists x. \text{man}'(x) \wedge Q(x) \rangle$$

Obviously, in a realistic model we should expect that the indefinite changes to an before vowel, that man can be modified by adjectives (and so the determiner can take not only bare nouns), and so on. However, these are matters of detail and do not bear on what we have to say in the sequel.

To say that **MAN** is a sign of English is to say that the string **man** if occurring as a syntactic object of category **N** has the meaning  $\lambda x. \text{man}'(x)$ . (which, by the way, is nothing else but **man'**). It is possible to have any number of signs with identical exponent, category or meaning. For example, the lexicon of English will contain at least two entries for **bank** as a common noun, one that has meaning roughly paraphrasable as ‘is a bank of a river’ and the other has meaning roughly paraphrasable as ‘is a financial institution’. The frameworks that come closest in spirit to this setup are Montague grammar and categorial grammar. However, as the exponents can be trees rather than strings, and even

complex functions on them, various other frameworks can be rendered into this form, ensuring that the approach we take is maximally neutral. However, in the course of this paper we shall make specific proposals as to how the categories shall look like and what operations other than  $\circ$  the grammar shall contain. How the requirements can factually be reconciled with particular frameworks, is another matter that lies outside the scope of this paper.

We assume in particular that the categories are attribute value structures, and that they contain a pair  $[\text{CASE}: \vec{\alpha}]$ . Here  $\vec{\alpha}$  is the **syntactic case** of the relevant element. We assume that cases are *sequences of morphemes* (formed with the help of ‘;’). Thus, there is no need for extra features to define cases. Roots have empty case. However, they may select items with a particular case.

As said, cases are sequences of morphemes, not just individual morphemes. A particular case in point, we argue, is constituted by the locatives. Morphologically, a locative is formed from a DP by the addition of two heads. This addition can proceed in two ways.

**Function Application** The meaning of the head is a function, and this function is applied to the meaning of the argument. Syntactically the operation performs slash–cancellation. This is the standard mechanism of categorial grammar, denoted by  $\circ$ .

**Case Stacking** The exponent  $e$  of the head is stacked as a case marker on the case stack. It replaces  $[\text{CASE}: \vec{\alpha}]$  by  $[\text{CASE}: \vec{\alpha} \wedge e]$ . Semantically, no change occurs. We denote this operation by  $\circledcirc$ .

We shall outline our analysis using the Finnish phrase *laivalta*, the ablative form of *laiva* (‘ship’). It is composed from three signs,

$$(7.5) \quad \text{LAIVA} := \langle \text{LAIVA}, DP[\text{CASE} : \varepsilon], \text{ship}' \rangle$$

$$(7.6) \quad \text{AT} := \langle 1, DP \setminus LP, \text{at}' \rangle$$

$$(7.7) \quad \text{COI} := \langle \text{ta}, LP \setminus MP, \text{from}' \rangle$$

It can mean three things:

(a) It can mean ‘from the ship’. In this case we say that it has **null syntactic case**. Its structure is

$$(\text{LAIVA} \circ \text{AT}) \circ \text{COI} = \langle \text{laivalta}, MP, \text{from}'(\text{at}'(\text{ship}')) \rangle$$

(b) It can mean ‘at the ship’. In this case we say that its **syntactic case** is the **cofinal**. Its structure is

$$(\text{LAIVA} \circ \text{AT}) \circledcirc \text{COI} = \langle \text{laivalta}, LP[\text{CASE} : \text{ta}], \text{at}'(\text{ship}') \rangle$$

(c) It can mean ‘the ship’. In this case we say that it has **ablative syntactic case**. Its structure is

(LAIVA@AT)@COF = ⟨laivalta, DP[CASE : 1; ta], ship'⟩

(Notice that 1; ta ≠ 1<sup>^</sup> ta = 1ta . The reason why we have to keep the two distinct is not apparent from the discussion of this paper.) The expression in (a) is an adverbial. It enters with its full meaning. Moreover, it is this meaning that motivates the case name ‘ablative’. (c) arises in case selection. For example, the verb *tuntua* selects ablative case:

(7.8) Tämä tuntuu laiva-lta/\*auto-sta.  
 this resembles ship-ABL/\*car-ELA  
*‘This looks like a ship/\*out of a car.’*

The reason why this is full case selection is that there is no choice: only ablative marked DPs can be used. Finally, the selection that gives rise to the meaning in (b) we call **directionality selection**. It occurs with verbs selecting only the directionality. It can be diagnosed by the fact that in place of the expression we can put in another one or a PP that has the same directionality.

(7.9) Jussi löysi raha-nsa laiva-lta/auto-sta.  
 Jussi found money-HIS ship-ABL/car-ELA  
*‘Jussi found his money on the ship/in the car.’*

Notice that in all three cases, the morphological realization is the same. Only the syntactic case and the meaning are different. Also, there is a competition between syntax and semantics: if the case is added as a syntactic marker, it is semantically void, and if the case enters with its proper meaning, then it cannot be stacked as a case marker. For more syntactic and semantic arguments in favour of this analysis see (kracht:against).

#### 4. Selection

Any PP can in principle also be selected by a verb. For example, German *Angst* *haben* (‘to be afraid’) selects *vor+DAT* (translated: ‘in front of’). (In (Kracht 2003), I argued that the selected case consists of three morphemes, not just two, as one might initially think.) In Hungarian, *félni* (‘to be afraid of’) selects ablative case, so it selects both M and L. In addition to these types of selection, there exist also the possibility of selecting just the M, not the entire case marker. This is directionality selection. Suppose for simplicity that the morpheme for cofinality in Finnish is -seen (in fact, this marker only appears after long vowels, but we do not intend to make things more complicated).

Then the verb *saapua* is the exponent of the following sign (which we contrast with the one for English):

(7.10) *SAAPUA* := ⟨*saapu*, *V/LP* [CASE: *een*], *arrive*⟩  
 (7.11) *ARRIVE* := ⟨*arrive*, *V/LP* [CASE: *ε*], *arrive*⟩

Contrast this with a verb that selects a case (that is, both *M* and *L*):

(7.12) *TUNTUA* := ⟨*tuntu*, *V/DP* [CASE: 1; *ta*], *resemble*⟩

Here is an example.

(7.13) *Saavuimme Lontoo*-seen.  
 arrived-we London-ILL  
 ‘*We arrived in London.*’

The case that must be used here is the illative (movement into). This has two reasons. (a) The state of being in a city is encoded using inner locative cases (except for a few Finnish places such as Turku), (b) the verb selects cofinal mode; hence, in place of the expected inessive (no movement), we find illative. To show that this is an instance of directionality selection and not ordinary case selection, we exchange Lontoo by *ranta* ‘coast’. Then allative case is mandatory.

(7.14) *Saavuimme rannalle*.  
 arrived-we coast-ALL  
 ‘*We arrived at the coast.*’

Notice that English does not tolerate cofinal mode. Neither does Finnish tolerate static mode.

(7.15) \**We arrived into London.*  
 (7.16) \**Saavuimme Lontoossa.*  
 arrived.we London-INE

Finnish has many verbs that are similar: *jäädä* ‘to stay, remain’, *unohtaa* ‘to forget’ (cofinal), *löytää* ‘to find’ (coinitial) (see (Fong, 1997), and other examples below). If *M* is a separate head, we expect that verbs which select only *M* will do so even with predicative and habitive cases. Moreover, the semantic contribution of *M* should be cancelled. We expect, for example, that the verb *jäädä* selects translative rather than inessive for predicatives, and allative rather than adessive for habitives. On the other hand, the verb *pysyä* ‘to remain’ selects static mode. Consequently, it chooses the essive, not the

translative (in the same meaning). (See (Fong, 2003) for an analysis along the lines of the earlier (Fong, 1997).)

(7.17) *Kuningatar jäi leske-ksi/\* leske-nä.*  
 queen remained widow-TRANS/\*widow-ESS  
*Kuningatar pysyi \*leske-ksi/leske-nä.*  
 queen remained \*widow-TRANS/widow-ESS  
*The queen remained a widow.*

Similarly, look at the following contrast with habitives:

(7.19) *Talo pysyi minulla.*  
 house remained me-ADE  
 (7.20) *Talo jäi minulle.*  
 house remained me-ALL

Notice that in our analysis *jäädä* selects not only *LPs* in cofinal mode, but also predicative phrases and habitives. Each of the different arguments has a different semantics, since the three are type-theoretically different. This is to be expected. Other verbs are not that flexible (for example *väsyä* ‘to get enough of, get tired’).

Hungarian enjoys selectional properties that are much closer to German than to Finnish. However, it also has verbs that select the cofinal, where the German (and English) counterparts select static mode. One example is *bújni* ‘to hide’. Another example is

(7.21) (Hu.) *Közel vagyunk a pályaudvar-hoz.*  
 close we.are the train.station-ALL  
*‘We are close to (sic!) the train station.’*

(Korhonen, 1996) claims that in Finno–Ugric languages the cofinal mode is the least marked one, while in Indo–European it is the static mode (see also (Alhoniemi, 1967)).

## 5. Significance for Interpretation

The primary difference between selected and unselected properties of a constituent is that the selected properties are semantically inert. For example, if Hu. *félni* selects a DP in ablative case, the ablative will not contribute to the meaning. This can be seen as a universal claim or just as a matter of coding. Surely, if a head selects an argument with such and such property (say, in cofinal mode), we can write whatever meaning this property contributes to the complex expression into the meaning of the head. However, if some property

(say, cofinality) is unselected, then its contribution is its normal one and there should be no need to encode it anywhere. To see the point here, notice that it is somehow reasonable that the cofinal appears with the Finnish verb *saapua*, as it is not necessarily logical that *ankommen* in German selects static mode. It is conceivable that there is a representation that makes this difference fall out. If we used such representations, however, we would implicitly claim that *saapua* means something different than *ankommen*. This would make translation next to impossible, though. The simplest approach is therefore to treat this as an instance of selection and give both verbs the same semantics.

The posture verbs are interesting in this connection. Some of them actually allow to use a directional (always the cofinal), in which case the verb denotes motion-to-posture. A case in point is provided by *Hu. 11ni*. When used with a static mode it is a posture verb (German *stehen*), while when used with a directional it is a verb of motion-to-posture (German *sich stellen*):

(7.22) *Romano Prodi[...]a Berlin-Párizsi vonal mellé áll.*  
 R. P. the Berlin-Paris line near-COF stand.  
*'Romano Prodi adopted the position of Berlin and Paris.'*

(Népszabadság Feb 13, 2003, commenting on the dispute between France, Germany and the USA.) In German, the distinction between posture and motion-to-posture is made lexically (see the example above and *sitzen* 'to sit' and *sich setzen* 'to sit down'). The contrast static/cofinal is actually signalled not by the preposition but by the dative/accusative contrast on the DP, as can be seen with pure motion verbs:

(7.23) *Sie liefen in den Wald.*  
 They ran in the-ACC forest  
 (7.24) *Sie liefen in dem Wald.*  
 They ran in the-DAT forest

With pure motion verbs, no difference in verb meaning arises, however. (The same contrast is coded in Mari (= Cheremiss) using the illative/lative contrast.) Some verbs in German can denote both posture and motion-to-posture without there being a visible difference. An example is *sich stützen auf* ('to rest on'). A similar verb is *sich verstecken* ('to hide'), which in contrast to Hungarian selects static. Thus, all four options are realized for motion and posture verbs:

- (1) The verb does not select mode. Different modes denote different paths of motion. (Example: motion verbs)

- (2) The verb does not select mode. Different modes induce different verbal meanings. (Example: posture and motion-to-posture contrast with Hu. **ällni**)
- (3) The verb selects static case. (Example: Ge. **sich verstecken**)
- (4) The verb selects cofinal case. (Example: Hu. **bújni**)

Case (2) could be analysed as involving two homophonous roots. However, the contrast is quite systematic so that this account would miss the general pattern.

## 6. Predicting Selectional Properties

Uralic languages are often very different from Indo-European languages as concerns the selection of *M*. For example, verbs of change of state often select coinitial or cofinal mode. (See (Fong 1997) and (Karlsson, 1984).)

- (7.25) (Fi.) **Rakennamme uuden hotellin Turkuun.**  
build-we new hotel Turku-ILLA  
'We are building a new hotel in Turku.'
- (7.26) (Fi.) **Ukko väsyi tie-lle.**  
old.man got.tired way-ALL  
'The old man got tired on (lit. onto) the road.'
- (7.27) (Fi.) **Joulu-na Jumala syntyi hevon heinähuonee-seen.**  
Christmas-ESS God was.born horse stable-ILL  
'At Christmas, God was born in (lit. into) a horse stable.'
- (7.28) (Fi.) **Somap' on sota-han kuolla.**  
sweet is war-ILL to.die  
'It is sweet to die in (lit. into) war.'
- (7.29) (Fi.) **Tää-ltä pyrkii häviämään tavarointa.**  
this-ABL tends disappear things  
'(From) here, things tend to disappear.'
- (7.30) (Fi.) **Metsästäjä ampui karhun metsä-än.**  
hunter shot bear forest-ILL  
'The hunter shot the bear in (lit. into) the forest.'
- (7.31) (Mari) **Wə·ðeško·ləšəwo·l'ək.**  
'The animal died in (lit. into) the water.'

The explanation according to Fong is as follows: the meaning of the verb has two phases (this is generally the case with verbs of creation, verbs of action,

and verbs of change of state). If the property holds at the end state, cofinal mode is used, if the property holds at the begin state, coinitial mode is used. To make this idea work, the directional meaning of *tielle* ('onto the way') and *metsään* ('into the forest') must be cancelled. Moreover, it would predict that a static locative is generally impossible — but the verb *pysyä* does select static mode.

Some of these examples could be dealt with inside a structured theory of the lexicon in the spirit of (Wechsler, 1995). We may postulate a generic verb of creation and coming to existence, which select cofinals for the location of its transitive object/subject. This solves (7.25) and (7.27) in a principled manner. Notice that static selection of English and German in these sentences might be deemed no less problematic, since the static mode seems to require the existence of the theme throughout the event time. Hence, for these verbs we have to allow for the fact that the object only exists at some subinterval. However, facts are complex; we have argued that static locatives predicate over the event location. It is only when we unfold the temporal patterns of these verbs that we see what this *actually* means. In the case of (7.25) we contend that the event of building takes place at a certain location inside Turku (the building site), which is independent of the existence of the building itself. Finnish employs a different metaphor: it considers the building a mover onto which the directional locative hooks. It predicates a change of location figuratively from somewhere into Turku. (An analogous analysis will work for (7.29), which does not require the existence of objects beyond their moment of disappearance.) However, if we wrote that into the meaning of the cofinal mode, there would be no principled way to stop it from overgeneralising. Hence, cofinality selection seems to be the best option.

A comparable case is that of coinitial locatives. We find here that Indo-European languages do use them more in line with Finnish (cf. (7.29)). They are also used in the meaning of 'location of source', for example with verbs of communication.

(7.32) **Er rief ihnen von einem Stein aus etwas zu.**

he shouted them-DAT from a stone PREP something to

(7.33) **Er zielte vom Hochsitz aus auf den Bären.**

he aimed from.the raised.hide PREP at the bear

(7.34) **Er rief seinen Anwalt von London aus an.**

he rang his lawyer from London PREP up

Notice that the circumposition *von+NP[DAT] aus* does not code the source (source is subject); rather, it encodes the location of the subject. If a plain inessive is used, that encodes either the location of the subject or that of the object (the position of the locative partially disambiguates, see (maiien-

born:modifiers)):

(7.35) **Er rief seinen Anwalt in London an.**  
 he rang his lawyer in London up (object/subject)

(7.36) **Er rief in London seinen Anwalt an.**  
 he rang in London his lawyer up (subject)

However, cofinal mode is impossible. Notice that with other places (die Bahamas ‘the Bahamas’), superessive replaces inessive (auf den Bahamas anrufen ‘to give a call in the Bahamas’), once again demonstrating that this is a case of directionality selection.

## 7. Mode Heads: Evidence from Mari

In (bierwisch:lexikon) it is assumed that there is selection of directionality, and that it is a matter of binary choice ( $[\pm \text{directional}]$ ). This would allow to save the account of single head selection, since now M and L are one head. Although languages have various modes, most of them are not grammaticalized (I know of no grammaticalisation of the recessive mode in German, for example). Mostly, the distinction between directional and nondirectional takes care of everything, particularly since the choice of the type of directional mode (coinitial/cofinal) seems to be predictable. Still, it seems that the best way is to assume that directionality selection is a case of head selection (which implies that it can have many more choices in principle). In an extensive study, (alhoniemi:wohin) has investigated the use and distribution of the lative and illative in Mari. Both are directional cases, and both the lative in the illative express cofinal mode. Alhoniemi notes that where a directional in Mari (and other Finno–Ugric languages) corresponds to a static locative in Indo-European, it is typically expressed by a lative (this is the case with the examples given above). For example, the place where someone undergoes change is expressed in the lative, quite unlike other Finno–Ugric languages. On the other hand, lative and illative sometimes are in free variation, sometimes not. Verbs of eating and drinking, for example, require a lative. There seems to be no theory in terms of the meaning that explains this. For such a theory would have to tell us which arguments may count as undergoing change; for these are the arguments that are predicated of using the lative. The choice lative versus other locative can only be predicted if we know independently which argument is changing. I know of no theory that can fulfill this. For notice that any argument in a verb that expresses a change undergoes change of some sort: its relation to the other arguments changes. For example, if I cook spaghetti, then not only the spaghetti change from uncooked to cooked, also I change: from someone standing in front of a pot of uncooked spaghetti into someone who does not. There is as far as I know no theory that defines the cut-off point

between the good cases and the bad ones. I conclude that there is every reason to believe that directionality selection is an instance of head selection, and that languages may have quite different sets of mode heads.

## 8. Conclusion

I have argued that locatives consist of minimally two parts, one specifying the mode, the other the place. There is, as far as I can see, no difference between the various realizations, be it by cases, be it by adpositions. Moreover, selection can take place either by selecting both heads or by just selecting one. Interestingly, the typical scenario of a PP selected by a head consists — under this analysis — of a selection of two, sometimes even three heads (see (kracht:against)). This is quite unlike what is assumed in current syntactic theories, where a head can only select the highest head inside its immediate complement. Interestingly, the case of selection of a single head does exist. This is what we call directionality selection. It has only rarely been studied. A proper understanding of its mechanics is however vital for many areas of linguistics and computational linguistics (we only mention machine translation, and man–machine interaction as cases in point).

## References

Alho Alhoniemi. *Über die Funktion des Wohin–Kasus im Tscheremissischen*. Number 142 in Suomalais–Ugrilaisen Seuran Toimituksia. Suomalais–Ugrilainen Seura, Helsinki, 1967.

Alho Alhoniemi. Suomen ja tšeremissin kielen suuntajärjestelmien funktionaalista rakenteesta. (The functional structure of the direction–case systems in Finnish and Cheremis.) Finnish with English summary. *Sananjalka*, 10:66 – 77, 1968.

Manfred Bierwisch. On the Grammar of Local Prepositions. In Manfred Bierwisch, Wolfgang Motsch, and Ilse Zimmermann, editors, *Syntax, Semantik und Lexikon*, Studia Grammatica XXIX, pages 1 – 65. Akademie Verlag, 1988.

Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors. *Language and Space*. MIT Press, Cambridge, Mass., 1996.

Vivienne Fong. *The Order of Things: What Directional Locatives Denote*. PhD thesis, Stanford University, 1997.

Vivienne Fong. Resultatives and Depictives in Finnish. In Diane Nelson and Satu Manninen, editors, *Generative Approaches to Finnic and Saami Linguistics*, pages 201 – 233. CSLI, Stanford, 2003.

Ray Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge, Massachusetts, 1983.

Fred Karlsson. *Finnische Grammatik*. Helmut Buske Verlag, Hamburg, 1984.

Mikko Korhonen. Merkmalhaftigkeit und Merkmallosigkeit in den finnisch-ugrischen Lokalkasussystemen. In Tapani Salminen, editor, *Typological and Historical Studies in Language by Mikko Korhonen*, number 223 in Suomalais-Ugrilaisen Seuran Toimituksia, pages 145 – 152. Suomalais-Ugrilainen Seura, Helsinki, 1996.

Marcus Kracht. On the semantics of locatives. *Linguistics and Philosophy*, 25:157 – 232, 2002.

Marcus Kracht. Against the Feature Bundle Theory of Case. In Ellen Brandner and Heike Zinsmeister, editors, *New Perspectives on Case*. CSLI, 2003.

Stephen Levinson. *Space in Language and Cognition. Explorations in Cognitive Diversity*. Number 5 in Language, Culture and Cognition. Cambridge University Press, Cambridge, 2003.

Sebastian Löbner. *Wahr neben Falsch*. Number 244 in Linguistische Arbeiten. Max Niemeyer Verlag, Tübingen, 1990.

Claudia Maienborn. On the Position and Interpretation of Locative Modifiers. *Natural Language Semantics*, 9:191 – 240, 2001.

Colin Masica. *The Indo-Aryan languages*. Cambridge Language Surveys. Cambridge University Press, 1991.

Igor Mel'čuk. *Cours de Morphologie Générale, 2ème Partie: Significations Morphologiques*. Les Presses Universitaires de Montréal, Montréal, 1994.

Seungho Nam. *The Semantics of Locative Prepositional Phrases in English*. PhD thesis, UCLA, 1995.

Soteria Svorou. *The Grammar of Space*. John Benjamins, Amsterdam, 1993.

Stephen Wechsler. *The Semantic Basis of Argument Structure*. CSLI, Stanford, 1995.

Dieter Wunderlich and M. Herweg. Lokale und Direktionale. In D. Wunderlich and A. von Stechow, editors, *Handbuch der Semantik*. Walter de Gruyter, 1986.

## Chapter 8

# VERB-PARTICLE CONSTRUCTIONS IN THE WORLD WIDE WEB

Aline Villavicencio

*Department of Language and Linguistics  
University of Essex  
Wivenhoe Park  
Colchester CO4 3SQ, United Kingdom  
avill@essex.ac.uk*

### Abstract

In this paper we investigate verb-particle constructions, discussing their characteristics and their availability for use with NLP systems. Combinations automatically extracted from corpora greatly improve the coverage of available lexical resources. However, the data sparseness problem is particularly acute for these constructions and even using a corpus as large as the British National Corpus, one finds that a great proportion of combinations have a very low frequency, while others never occur in it. To minimise the problem of data sparseness in this paper we propose to validate candidate VPCs using the World Wide Web as a very large corpus. This method can be used to extend the coverage of existing lexical resources by filtering combinations automatically generated from classes of verbs, and by improving the reliability of those combinations automatically extracted from corpora.

**Keywords:** Verb-Particle Constructions, Verbal Classes, World Wide Web, Productivity.

### 1. Introduction

In this paper we investigate verb-particle constructions (VPCs) in English and their availability for NLP systems. Due to their complex characteristics and their flexible nature, they provide a challenge for NLP technology. In par-

ticular, there is a lack of adequate resources to identify and treat VPCs, and many applications cannot capture them appropriately. However, due to their frequency in natural language interactions, it is clear that successful applications need to deal with them adequately, if they propose to capture natural languages successfully.

VPCs are combinations of verbs and prepositional or adverbial particles, such as *eat up* in *Bob ate up the chocolate*. In these constructions particles are characterised by containing features of motion-through-location and of completion or result in their core meaning (Bolinger, 1971). However, VPCs can range from these more regular combinations, such as *clean up* (in e.g. *He needs to clean up his flat*) to idiosyncratic or semi-idiosyncratic ones, such as *give in* (in e.g. *Her son was so determined to get what he wanted that she finally gave in*). Cases of ‘idiomatic’ VPCs like *give in*, meaning *to agree to what someone wants after a period when you refuse to agree*, where the meaning of the combination cannot be straightforwardly inferred from the meaning of the verb and the particle, fortunately seem to be a small minority (Side, 1990). Most cases seem to be more regular, with the particle compositionally adding a specific meaning to the construction and following a productive pattern. Indeed, Side noted that particles in VPCs seem to fall into a set of possible categories, defined according to their meanings in the combinations. For instance, in his analysis of VPCs involving *off*, which is defined as *indicating distance in time or space, departure, removal, disconnection, separation*, most VPCs considered seem to fit into this category. Examples are *take off* meaning to depart, *cut off* meaning to disconnect and *strain off* to remove. A three way classification is adopted by Dehé (Dehé, 2002), Emonds (Emonds, 1985) and Jackendoff (Jackendoff, 2002), where a VPC can be classified into compositional, idiomatic or aspectual, depending on its sense. In the compositional VPCs the meaning of the construction is determined by the literal interpretations of the particle and the verb (e.g. *throw out* in *I don't want these old books anymore, so I'll throw them out*). Idiomatic VPCs, on the other hand, cannot have their meaning determined by interpreting their components literally (e.g. *go off* meaning ‘to explode’ in *Dur! ing the last war a bomb went off near that village*). The third class, of aspectual VPCs, have the particle providing the verb with an endpoint, suggesting that the action described by the verb is performed completely, thoroughly or continuously (e.g. *tear up* in *She'll tear up any letters that he sends her*). In the investigation described here the focus is on compositional and aspectual senses of combinations of verbs and particles.

VPCs have been the subject of a considerable amount of interest, and some investigation has been done on the subject of productive VPCs. Bame analysed some of these productive cases in the framework of Head-Driven Phrase Structure Grammar: namely those of aspectual and resultative combinations using the particle *up* (Bame, 1999). For example in *Kim carried the television up* the

resultative *up* indicates that the argument is affected (i.e., at the end of the action the television is *up*). In contrast, the aspectual *up* in *Kim ate the sandwich up* suggests that the action is taken to some conclusion (i.e., the sandwich is totally consumed at the end of the action). Villavicencio and Copestake proposed defining a family of lexical rules, organised in a default inheritance hierarchy, to capture productive patterns of verb-particle constructions like these (Villavicencio and Copestake, 2002). Fraser pointed out that semantic properties of verbs can affect their possibilities of combining with particles (Fraser, 1976). For example *bolt*, *cement*, *clam*, *glue*, *paste* and *nail* are all semantically similar verbs where the objects specified by the verbs are used to join material and they can all productively combine with *down*. There is clearly a common semantic thread running through this list, so that a new verb that is semantically similar to them can also be reasonably assumed to combine with *down*. Moreover, Side notes that frequently new VPCs are formed by analogy with existing ones, with often the verb being varied and the particle remaining (e.g. *hang on*, *hold on* and *wait on*).

As these works suggest, many VPCs follow productive patterns, where semantically related verbs are combined with a given sense of a particle. By identifying classes of verbs that follow patterns such as these in VPCs, it is also possible to maximise the use of the information contained in lexical resources. In this way, one can make use of regular patterns to productively generate VPCs from verbs already listed in a lexical resource, according to their verbal classes and the particles with which they can combine. For example, the resultative combinations *walk/run/jump up/down/out/in/away/around* from the motion verbs *walk*, *run* and *jump* and the directional/locative particles *up*, *down*, *out*, *in*, *away* and *around*. In this context, the use of Levin's classification of verbs (Levin, 1993) to productively generate candidate VPCs from semantically related verbs is a possible alternative to extend the coverage of lexical resources, as suggested by Villavicencio (Villavicencio, 2003). The verbal classes seem to be good indicators of productivity in verb-particle constructions. However, the data sparseness problem, which is particularly acute for multiword expressions like VPCs, means that the full contribution made by the candidate VPCs remains yet! to be determined, since a large part of the combinations proposed could not be verified given the available corpus and lexica. From these combinations some may be valid, but simply do not occur in these resources, while others are genuinely invalid. In this paper we propose to verify the validity of VPCs automatically generated from classes of verbs by searching for them using the World Wide Web as a very large corpus, in order to minimise the problem of data sparseness, following Grefenstette (Grefenstette, 1999) and Keller et al. (Keller et al., 2002).

We begin by discussing some characteristics of VPCs that make them so challenging. Then in the next two sections we analyse the coverage provided

by some available lexical resources, and the use of information automatically extracted from corpora to extend their coverage. We then discuss Levin's classes of verbs and the combinations they productively generate with the particle *up*, which is the most widely used particle in these lexical resources. Next we address the issue of how these can be validated using the World Wide Web, to avoid the problem of data sparseness, closing with a discussion of the results obtained and future work.

## 2. VPCs in a Nutshell

In this section we briefly discuss some of the characteristics that make VPCs so challenging for NLP.<sup>1</sup> VPCs are often highly polysemous, with, for instance, eight senses being listed for *make up* in the Collins Cobuild Dictionary of Phrasal Verbs (among them, e.g. *to form something* and *to invent*). They also show syntactic variation, where each combination can take part in several different subcategorisation frames. For example, *add up* can occur as an intransitive verb-particle combination in *It's a few calories here and there, and it all quickly adds up* or as a transitive one in *We need to add these marks up*.

In transitive VPCs, where an NP complement is required, some particles have a fixed position in relation to the verb, such as *come up* in *She came up with the idea*, where the particle is expected immediately after the verb. Thus one cannot have *\*She came with the idea up*. Other combinations have a more flexible order in relation to the verb, and can equally well occur after another complement or immediately after the verb: e.g. *John ate his cereal up* and *John ate up his cereal*. In the latter, the particle comes before a simple definite NP without taking it as its object (unlike, e.g., *It consists of two parts*, which is a prepositional verb). Whether a particle can be separated or not from the verb may depend on the degree of bonding between the particle and the verb, on the size of the NP, and on the kind of NP. Thus, when the NP is an unstressed personal pronoun, in a transitive VPC, it must precede the particle (e.g. *They ate it up* but not *\*They ate up it*). This is also the case for VPCs subcategorising for other verbal complements, like PPs and sentential complements, ! where the particle must come immediately after the verb (e.g. *He found out about the affair* but not *\*He found about the affair out*). Besides complements, certain adverbs are also accepted between the verb and the particle, such as *right in* in *He came right back*.

## 3. VPCs and Dictionaries

In this section we analyse some of the lexical resources available for NLP systems, in terms of the VPCs they contain. Table 8.1 shows the coverage of phrasal verbs (PVs) in several dictionaries and lexica: Collins Cobuild Dictionary of Phrasal Verbs (Collins-PV), Cambridge International Dictionary of

Phrasal Verbs (CIDE-PV), the electronic versions of the Alvey Natural Language Tools (ANLT) lexicon (Carroll and Grover, 1989) (which was derived from the Longman Dictionary of Contemporary English, LDOCE), the COMLEX lexicon (Macleod and Grishman, 1998), and the LinGO English Resource Grammar (ERG) (Copestake and Flickinger, 2000) version of November 2001. This table shows in the second column the number of PV entries for each of these dictionaries, including not only verb-particle constructions but also prepositional verbs. The third column shows the number of VPC entries (available only for the electronic dictionaries).

Table 8.1. Phrasal Verb Entries in Dictionaries

Dictionary	PV Entries	VPC Entries
ANLT	6,439	2,906
CIDE-PV	over 4,500	-
Collins-PV	over 3,000	-
Comlex	12,564	4,039
ERG	533	337

These dictionaries have a considerable number of PV entries potentially providing us with a good starting point for handling VPCs. Each dictionary uses a slightly different set of verbs and particles in its VPCs, and table 8.2 shows some of their characteristics. In this table A+C represents the union of ANLT and Comlex, A $\cap$ C their intersection and A+C+E the union of ANLT, Comlex and ERG.

Table 8.2. VPCs in Dictionaries

Dictionary	Verbs	VPCs Entries	Distinct VPCs	Particles	Verbs in VPCs
ANLT	5,667	2,906	2,250	44	1,135
Comlex	5,577	4,039	1,909	23	990
ERG	1,223	337	270	25	176
A+C	6,043	-	3,107	44	1,394
A $\cap$ C	5,201	-	1,052	23	731
A+C+E	6,113	-	3,156	45	1,400

When the particles were ranked according to the frequency with which they occur in the VPCs, similar patterns were obtained for all of the dictionaries. Figure 8.1 shows the five top ranked particles for each of the dictionaries. For all of them, *up* is the particle involved in the largest number of combinations.

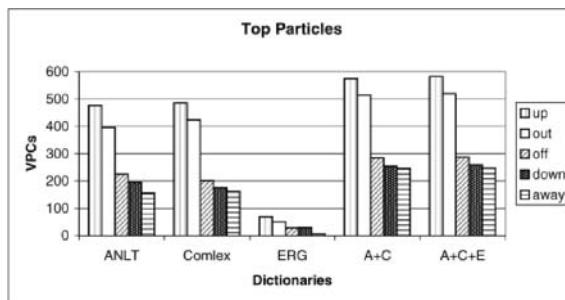


Figure 8.1. Top Ranked Particles in Dictionaries

In each of these dictionaries only a small proportion of the total number of verbs is used in VPCs, as can be seen in figure 8.2 which shows the proportion of verbs used in VPCs from all the verbs listed in a dictionary. For example, only 20% of the verbs listed in the ANLT form at least one VPC. For the other dictionaries this proportion is even lower. These tend to be very widely used and general verbs, such as *come*, *go*, *get*, *put*, *bring* and *take*. Which of the remaining verbs do not form valid VPCs and which verbs form VPCs that were simply omitted remains to be determined, and this investigation is an attempt to take a step in this direction.

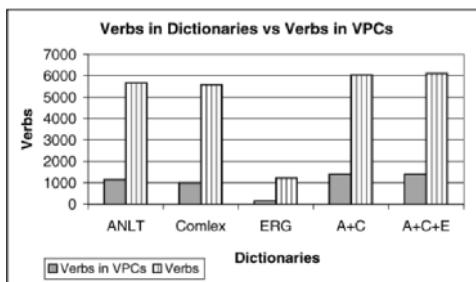


Figure 8.2. Verbs in VPCs and Verbs in Dictionaries

The number of VPCs listed in each dictionary is shown in table 8.2, where we can also see the increase in the number of VPCs obtained by the union of the dictionaries. Even though there is a large number of entries already obtained by combining the two largest dictionaries, ANLT and Comlex, a considerable proportion (16%) of the entries in the LinGO ERG lexicon are not listed in any of them (this proportion would increase if subcategorisation etc was also taken into account).<sup>2</sup> Most of these are at least semi-compositional, e.g. *crisp up*, *come together*, *tie on*, and were probably omitted from the dic-

tionaries for that reason,<sup>3</sup> though some others, such as *hack up*, are probably recent coinages. Dictionaries are valuable but static resources that tend to list idiosyncratic combinations at the expense of omitting the more productive ones, so one cannot rely only on the combinations they provide.

#### 4. VPCs and Corpora

The number of VPCs available in language is constantly growing, and ways of extending the coverage provided by lexical resources are needed. The use of corpora to extract VPCs is a good alternative for extending the coverage of these resources. In this section we use VPCs extracted from the British National Corpus (BNC) and compare them with those contained in the combined dictionaries (A+C+E-VPCs), using the former to complement the coverage provided by the latter.

The BNC (Burnard, 2000) is a 100 million word corpus containing samples of written text from a wide variety of sources, designed to represent as wide a range of modern British English as possible. It includes texts from newspapers, journals, books, and many other sources. Using the methods proposed by Baldwin and Villavicencio (Baldwin and Villavicencio, 2002), 8,751 VPC entries were extracted from the BNC. These entries are classified into intransitive and/or transitive VPCs, depending on their subcategorisation frame, and they result in 7,078 distinct VPCs. A few of these entries are not VPCs but rather noise, such as \*\*\*\* *off's down*, etc. After removing the most obvious cases of noise, there were 7,070 VPCs left. These are formed by 2,542 verbs and 48 particles. The method proposed by McCarthy et al. (McCarthy et al., 2003) resulted in 4,482 distinct VPCs extracted, after the most obvious cases of noise were removed. They are formed by the combination of 1,999 verbs and 9 particles, among which there are also cases of prepositional verbs.

These different extraction methods yielded different sets of VPCs, as it is possible to see in table 8.3. This table shows some comparisons, where BNC-1 represents the set of VPCs extracted using the methods described by Baldwin and Villavicencio, BNC-2 those extracted by McCarthy et al., and BNC the union of both. Even though these two methods were applied to the same corpus, their results are quite distinct, with one complementing the other.

In terms of the VPCs, by joining A+C+E-VPCs with all the VPCs extracted from the BNC (BNC-VPCs) there is an increase of 209% in the number of VPCs, since from the 8,911 VPCs in BNC, only 2,318 are also in the combined dictionaries, as can be seen in table 8.4. A considerable number of the extracted VPCs form productive combinations, some containing more informal or recent uses of verbs (e.g. *hop off*, *kangaroo down* and *skateboard away*). These VPCs provide a useful addition to the information contained in the dictionaries, resulting in a total of 9,745 distinct combinations.

Table 8.3. Comparison between VPCs Automatically Extracted from Corpora

Resources	VPC Entries	Verbs	Particles
BNC-1	7,070	2,542	48
BNC-2	4,482	1,999	9
BNC-1 - BNC-2	4,429	956	39
BNC-2 - BNC-1	1,841	413	0
BNC-1 $\cap$ BNC-2	2,641	1,586	9
BNC	8,911	2,955	48

Table 8.4. Comparison between VPCs from Combined Dictionaries and those from BNC

Resources	VPC Entries	Verbs	Particles
A+C+E	3,156	1,400	45
BNC	8,911	2,955	48
A+C+E - BNC	834	160	17
BNC - A+C+E	6,593	1,715	20
A+C+E $\cap$ BNC	2,318	1,240	28
A+C+E+BNC	9,745	3,115	65

These methods provide us with a larger set of VPCs and some information about their syntactic behaviour, like their subcategorisation frames. However, they suffer from the problem of data sparseness and a great proportion of the extracted VPCs have a very low frequency. For instance, in BNC-2 40.52% of the combinations occur only once. Among these there are genuine combinations (e.g. *telephone back*) but there are also instances of false positives like misspellings or noise (e.g. *scimitare down* instead of *scimitar down* and *theyre in* instead of *they're in*), and it is difficult to decide which is which on the basis of one occurrence. In the next section we discuss a possible way of using the World Wide Web to help distinguish genuine VPCs.

## 5. VPCs in the Web

One possible way of minimising the problem of data sparseness is to use the World Wide Web as an extremely large corpus, since, as pointed out by Grefenstette (Grefenstette, 1999) and Keller et al. (Keller et al., 2002), the web is the largest data set available for NLP: in December 2002 the web contained at least 3,033 million pages, which were indexed by the search engine Google, according to the Search Engine Showdown (<http://www.searchengineshowdown.com>). Several researchers have started to

explore this idea, making use of this huge resource to overcome the problem of data sparseness. For instance, Grefenstette employs the web to do example-based machine translation of compounds from French into English. The method he employs would suffer considerably from data sparseness if it were to rely only on corpus data, so for compounds that are sparse in the BNC he also obtains frequencies from the web. Keller et al. use the web to obtain frequencies for adjective-noun, noun-noun and verb-object bigrams, testing if the web could be used to obtain frequencies for bigrams that are unseen in a given corpus. They suggest that the large amount of data available on the web largely outweighs any problem that may derive from it being unbalanced and containing noise.

In this paper we propose to use the web to find evidence to distinguish between valid VPCs and noise in automatically generated or extracted combinations, minimising the problem of data sparseness. The web could be thus employed to obtain frequencies for the candidate VPCs and filter them out accordingly. To test these possibilities, initially all VPCs in BNC-2 that have frequency of 1 occurrence were searched on the web using the search engine Google. For each combination searched, Google provided us with a measure of frequency in the form of the number of pages in which that combination appeared. Indeed, as expected the results obtained indicate that this alternative provides further evidence for differentiating spurious combinations such as *scimitare down*, which cannot be found, or are found in only a small number of pages, from genuine VPCs like *package up* (in e.g. *Why do I need to use a zip program to package up my files?*).

In order to investigate even further the contribution of the web, a verbal classification was used to automatically generate candidate VPCs, and the web used as a corpus to test the validity of the combinations generated. In this investigation we concentrate on VPCs generated by combining a classification of semantically related verbs and the particle *up*. The valid combinations can then be used to extend the coverage of the available resources.

## 5.1 The Candidate VPC Set

Fraser noted how semantic properties of verbs can affect their possibilities of combination with particles (Fraser, 1976). For example verbs of hunting and the resultative *down* (*hunt/track/trail/follow down*) and verbs of cooking and the aspectual *up* (*bake/cook/fry/broil up*). Therefore, by having a semantic classification of verbs one can investigate how they combine with certain particles. This can be used to extend the coverage of the available resources by generating VPCs from classes of related verbs that follow productive patterns of combinations. One such classification was proposed by Levin, where verbs are grouped into classes according to semantic and syntactic properties,

based on the assumption that the syntactic behaviour of verbs is semantically determined (Levin, 1993). In this section we further investigate the possibility of using Levin's classes of verbs to generate candidate verb-particle combinations, following Villavicencio (Villavicencio, 2003).

In Levin's classification there are 190 classes and subclasses that capture 3,100 different verbs, resulting in 4,167 entries, since each verb can belong to more than one class. For example, the verb *to run* belongs to classes 26.3 (Verbs of Preparing), 47.5.1 (Swarm Verbs), 47.7 (Meander Verbs) and 51.3.2 (Run Verbs). The number of elements in each class varies considerably, so that 60% of all of these classes have more than 10 elements, accounting for 88% of the verbs, while the other 40% of the classes have 10 or less elements, capturing the remaining 22% of the verbs. The 5 larger classes are shown in table 8.5.

Table 8.5. Five Larger Classes

Class	Class Name	Entries
45.4	Other alternating verbs of change of state	257
31.1	Amuse	220
51.3.2	Run	124
43.2	Sound emission	119
9.9	Butter	109

All the combinations formed by Levin's classes and the particle *up* were produced. The combinations were generated by taking each verb and appending the particle to it. It is necessary to test the validity of a candidate VPC, since not all verbs can be combined with particles. For example, Fraser noted the generalisation that stative verbs almost never combine with a particle (e.g. *know*, *want*, *hope*, *resemble*, etc); some other verbs seem to occur with only one particle (e.g. *chicken out* and *sober up*) (Fraser, 1976). Moreover, although there are some cases where it appears reasonable to treat verb-particle combination as fully productive (within fairly finely specified classes), there are also cases of semi-productivity. For instance, many verbs denoting cooking processes can occur with aspectual *up*: e.g. *boil up*, *fry up*, *brew up*, *heat up*. But some other combinations seem odd e.g. *?sauté up*. This problem of semi-productivity is further discussed by Villavicencio and Copestake (Villavicencio and Copestake, 2002). Nonetheless, some verbal classes (and particles) seem to be good indicators of VPC acceptability. For example, in Class 11.3 (Verbs of Bring and Take), all verbs seem to form valid combinations with the particles *in*, *down*, *out*, *up* (Villavicencio, 2003).

## 5.2 Looking for VPCs in the Web

From the 4,167 verbs listed in Levin's classification the majority, 3,933, are in the combined resources. However, from the 4,167 possible VPCs generated from combining the verbs in Levin's classes with *up*, only 1,674 are in the combined resources (A+C+E+BNC-VPCs). Even though the combined resources have a large number of VPCs, this coverage is still limited. For instance, in a manual analysis of the combinations involving the class of motion verbs, a great proportion of the VPCs are not attested in these resources, even if most of the combinations are considered acceptable by native speakers. It is necessary to establish whether the unattested VPCs genuinely do not form valid combinations, or whether they do not occur due to the data sparseness problem. In this section we discuss how to use the web to verify if the candidate VPCs are genuine on the basis of their occurrences on the web.

As not all verbs in Levin's classes will form valid VPCs, each of the combinations that was unattested in the combined resources was searched on the web using Google, which returned the number of pages in which that combination appeared. Since the goal is to be able to identify genuine cases, we assume that if a VPC is attested either in the combined resources or in the web, then it is a valid VPC.

In order to provide a uniform search pattern for all the VPCs, initially they were all searched as intransitive VPCs, which is one of the most common sub-categorisation frames for VPCs. Furthermore, it was necessary to define delimiters to use when searching for VPCs to ensure that *up* is not followed by an NP, which would be ambiguous between a transitive VPC (Verb Particle NP) and a prepositional verb, where the PP is headed by *up* (Verb PP), aiming to retrieve only VPCs, and not prepositional verbs. In this way, the following pattern was used for the searches: “⟨VERB⟩ up ⟨DELIMITER⟩”, where each verb in Levin's classes is searched for occurrences where it is immediately followed by “*up*” and a delimiter. Prepositions seem to be suitable candidates for delimiters of VPCs, and in this investigation, the prepositions *for* and *from* were used as delimiters. For instance, *slim up*, which was unattested in the combined resources, was found in the web in the context of *slim up for*, and one of the pages found contains the sentence: *Why do we need to spend tax money to convince you to slim up for your own good?*.

By adding a delimiter as an extra term in the context of the search, the aim is to avoid the problem of ambiguity with prepositional verbs. However, at the same time the addition of a delimiter also restricts considerably the evidence that can be gathered for the validity of a VPC, because any additional word in the search term may reduce the number of pages that can be retrieved. For example, searching only for *slim up* returns 1,400 pages against 42 returned by searching for *slim up for*. In this way, one exchanges the retrieval of a

potentially larger set of pages that contains the VPC for a much smaller but much more precise set containing the delimiter preposition too. Only pages containing that exact search pattern are retrieved.

An analysis of the results obtained confirms that *for* is a good delimiter of VPCs, since it is frequent enough in the data to allow for a considerable number of pages to be retrieved for a large number of combinations. Nevertheless, for greater accuracy other delimiters also need to be employed, since some of the unattested VPCs may occur so unfrequently that only using their occurrences with a single delimiter such as *for* as evidence may be too restrictive. In this way, when *from* is used as an alternative delimiter, a considerable increase in the number of attested VPCs was observed.

When looking for evidence to validate the candidate VPCs it is also important to consider that some of these combinations may be realised predominantly in a certain subcategorisation frame and therefore using only one frame in the search patterns may prove to be insufficient for validating them. Thus, for greater accuracy, alternative search patterns, with different subcategorisation frames may also be employed. For the purposes of this investigation, to also gather evidence for transitive VPCs, the pattern “⟨VERB⟩ ⟨PRON⟩ up ⟨DELIMITER⟩” was used, where PRON stands for the pronouns like *you*, *it* and *them*. The NP complement is in the form of personal pronouns added to the search pattern, since the use of pronouns not only simplifies the form of the NP, but also abstracts away from the problem of the VPC word order, given that pronouns tend to occur between the verb and the particle in transitive VPCs. These alternative subcategorisation frame patterns can also be useful to discover which of the frames is the preferred form for the realisation of a VPC, in the case of VPCs that occur in several different frames. An example is *eat up*, that can be used both as an intransitive VPC (e.g. in *let it eat up for a few hours*) or as a transitive one (e.g. *he should eat that up for breakfast*), but that is found in 141 pages as an intransitive VPC against only 4 pages as a transitive one.

In this investigation the results reported include both the intransitive and transitive search patterns and both *for* and *from* as delimiters.

Using the web as a corpus, a total of 1,871 of the candidate VPCs were considered valid. Among the unattested combinations one can find *genuflect up* and *salaam up*. As a result a total of 3,225 VPCs out of the 4,167 candidate VPCs was attested in the combined resources or in the web, corresponding to 77.4% of the possible candidates. From these, 154 are cases of VPCs containing verbs that were listed in the combined resources but were not used in any VPC listed therein. In terms of the classes, 96.31% of them had most of its candidate VPCs considered valid; from the remaining 3.68% of the classes, only one of them had no attested VPCs: Class 39.4 of Devour Verbs, which contain a total of 5 VPCs. By joining them with A+C+E+BNC-VPCs there is

an increase of 12.8% in the number of VPCs with a total of 10,994 VPCs. This is an encouraging outcome for this investigation, with the use of these very simple search patterns and with no extra linguistic processing required.

## 6. Conclusions

In this paper, we investigated using the web as a way of verifying the validity of candidate VPCs automatically generated from a verbal classification. This method is employed to find evidence for genuine VPCs, where we assume that a VPC is valid if it is attested, so that prior to inclusion in a lexical resource, unattested cases can be filtered out. However, one could also set a threshold for VPCs that are only attested in the web. In this case, adopting a threshold of e.g. 5 pages, a VPC would only be considered valid if it occurred at least in 5 pages. By searching the web for candidate VPCs generated from Levin's classes on the basis of their semantic/syntactic interrelations rather than searching for any possible occurrence of a word followed by a particle, means that only genuine verbs were used in the combinations avoiding random noise caused by misspelled words, non-native speakers, pages in other languages, etc. Thus, Levin's classes were used as a means of constraining the possible combinations and the web as a means of filtering unattested VPCs. However, as some of the verbs in Levin's classes can also occur as nouns (e.g. *mail*), for greater accuracy in the number of pages obtained for these cases, one alternative is to use the inflected forms of these verbs in the search patterns (e.g. *mailed up for, to mail up for,...*). This method was used to help not only to validate automatically generated VPCs, but also to verify which of the low frequency VPCs automatically extracted from corpora are genuine ones, while at the same time reinforcing their frequencies using the web. The valid combinations can then be used to extend the coverage of lexical resources.

The results obtained suggest that Levin's classes are indeed a good starting point for obtaining productive patterns in verb-particle constructions. This investigation focused only on the particle *up* as a test case, but it is already possible to see an improvement in the coverage of the available lexical resources when VPCs with this particle are considered. A more wide investigation using a larger set of verbs and particles and human annotators is envisaged, to extend even further the coverage of existing lexical resources. In particular, we plan to explore the use of other verbal taxonomies (e.g. (Tenny, 1995)) for generating VPCs, since some of them may prove to be even more suitable for this task than Levin's, given that the latter was not designed especially on the basis of VPCs. This investigation will continue to address the question of the great number of the verbal entries in a lexical resource not used in its VPCs, using the web to search for candidate VPCs generated by these verbs. For these automatically generated VPCs, a method like that proposed by Bannard and Baldwin (Ban-

nard and Baldwin, 2003), Baldwin (Baldwin, this volume) or McCarthy et al. (McCarthy et al., 2003) can be subsequently used to determine the semantics of the valid VPCs.

The approach proposed here can be straightforwardly extended to also deal with cases of combinations between other classes of words. For instance, it could be used to search for evidence of combinations of prepositions and nouns to aid e.g. the analysis done by McMichael (McMichael, this volume).

The results obtained so far are encouraging and confirm that the coverage of lexical resources can be straightforwardly extended by using (semantic) classifications of verbs to productively generate possible VPCs, and the web as a very large corpus to validate them.

## Acknowledgments

I'd like to thank Timothy Baldwin and Diana McCarthy for kindly making available the combinations extracted from the BNC. My thanks also to Ann Copestake and Marina Terkourafi for their comments. This research was supported in part by the NTT/Stanford Research Collaboration, research project on multiword expressions.

## Notes

1. A more detailed discussion of VPCs can be found in (Bolinger, 1971), (Fraser, 1976) and (Villavicencio and Copestake, 2002) among others.
2. The LinGO ERG lexicon was manually constructed with most of the verb-particle entries being empirically motivated by the Verbmobil corpus. It is thus probably reasonably representative of a moderate-size domain-specific lexicon.
3. The Cobuild Dictionary explicitly states that literal meanings and combinations are not given for all verbs.

## References

Bame, Ken. Aspectual and Resultative Verb-Particle Constructions with Up. *Handout for talk presented at the Ohio State University Linguistics Graduate Student Colloquium*, 1999.

Baldwin, Timothy and Villavicencio, Aline. Extracting the Unextractable: A Case Study on Verb-Particles. *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan , 2002.

Bannard, Colin and Baldwin, Timothy. Distributional Similarity and Preposition Semantics. *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France, 2003.

Bolinger, Dwight. *The Phrasal Verb in English*. Harvard: Harvard University Press, 1971.

Burnard, Lou. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services, 2000.

Carroll, John and Grover, Claire. The Derivation of a Large Computational Lexicon of English from LDOCE. In Branimir Boguraev and Ted Briscoe (eds.) *Computational Lexicography for Natural Language Processing*, Longman, 1989.

Copestake, Ann and Flickinger, Dan. An Open-Source Grammar Development Environment and Broad-Coverage English Grammar Using HPSG. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Las Palmas, Canary Islands, 2000.

Dehé, Nicole. *Particle Verbs in English: Syntax, Information Structure and Intonation*, Amsterdam/Philadelphia:John Benjamins, 2002.

Emonds, Joseph E. *A Unified Theory of Syntactic Categories*, Dordrecht:Foris, 1985.

Fraser, Bruce. *The Verb-Particle Combination in English*, New York:Academic Press, 1976.

Grefenstette, Gregory. The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *Proceedings of ASLIB, Conference on Translating and the Computer*, London, England, 1999.

Jackendoff, Ray. English Particle Constructions, the Lexicon, and the Autonomy of Syntax. In Nicole Dehé, Ray Jackendoff, Andrew McIntyre and Silke Urban (eds.) *Verb-Particle Explorations*, Berlin: Mouton de Gruyter, 2002.

Keller, Frank, Lapata, Maria, and Ourioupina, Olga. Using the Web to Overcome Data Sparseness. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, USA, 2002.

Levin, Beth. *English Verb Classes and Alternations - A Preliminary Investigation*, The University of Chicago Press, 1993.

Macleod, Catherine and Grishman, Ralph. *COMLEX Syntax Reference Manual, Proteus Project*, 1998.

McCarthy, Diana, Keller, Bill and Carroll, John. Detecting a Continuum of Compositionality in Phrasal Verbs. In Francis Bond and Anna Korhonen and Diana McCarthy and Aline Villavicencio (eds.) *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, 2003.

Side, Richard. Phrasal Verbs: Sorting Them Out. *ELT Journal*, 44(2):144–52, 1990.

Tenny, Carol. How Motion Verbs are Special: The Interaction of Semantic and Pragmatic Information in Aspectual Verb Meanings. *Pragmatics and Cognition*, 3(1): 31–73, 1995.

Villavicencio, Aline and Copestake, Ann. Verb-Particle Constructions in a Computational Grammar of English. In Jongbok Kim and Stephen Wechsler

(eds.) *Proceedings of the Ninth International Conference on Head-Driven Phrase Structure Grammar*, Seoul, Korea, 2002.

Villavicencio, Aline. Verb-Particle Constructions and Lexical Resources. In Francis Bond and Anna Korhonen and Diana McCarthy and Aline Villavicencio (eds.) *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, 2003.

## Chapter 9

# PREPOSITIONAL ARGUMENTS IN A MULTILINGUAL CONTEXT

Valia Kordoni

*Dept. of Computational Linguistics, Saarland University  
P.O. Box 15 11 50, D-66117, Saarbruecken Germany*  
kordoni@coli.uni-sb.de

**Abstract** This paper focuses on indirect prepositional arguments in a multilingual context. We show that the theoretical framework of HPSG (Pollard and Sag, 1994) with semantic representations in Minimal Recursion Semantics (MRS; (Copestake et al., 1999); (Copestake et al., 2001)) constitutes the appropriate theoretical basis for a robust, linguistically-motivated account of indirect prepositional arguments. The case study here is indirect prepositional arguments in Modern Greek and English.

**Keywords:** HPSG, MRS, robust deep analysis of indirect prepositional arguments, multilingual grammar development.

## 1. Introduction

This paper focuses on the key role of semantics in a robust deep analysis of indirect prepositional arguments in a multilingual context.

The aim here is to show that the theoretical framework of Head-Driven Phrase Structure Grammar (HPSG; (Pollard and Sag, 1994)) with semantic representations in Minimal Recursion Semantics (MRS; (Copestake et al., 1999); (Copestake et al., 2001)) constitutes the appropriate theoretical basis for a robust, linguistically-motivated account of indirect prepositional arguments, which does not only overcome the natural limitations of previous syntactic and semantic analyses of these arguments (see, among others, (Rappaport and Levin, 1988), (Pinker, 1989), (Markantonatou and Sadler, 1996)), but also provides the necessary formal generalizations for the analysis of such arguments

in a multilingual context, since MRS structures are easily comparable across languages. Our case study here is indirect prepositional arguments in Modern Greek and English (see Section 1.2).

The rest of the paper is organized as follows. In the next section (Section 1.2) we give an overview of the relevant data in Modern Greek and English. In Section 1.3 we present briefly previous analyses of indirect arguments. In Section 1.4 we present the robust analysis of indirect prepositional arguments that we propose. Finally, in Section 1.5 we are highlighting the advantages the proposed analysis presented in Section 1.4 might bring to the task of (multilingual) development of broad coverage grammars of natural language.

## 2. The Data

In this section we turn to the data. The given English data is extensively discussed in the existing literature on valency patterns and valency alternations, while the Modern Greek is not. Our overview focuses on whether the optional arguments of a verbal predicate are existentially quantified in their absence. From this point of view, a three-way classification emerges as syntactically obligatory arguments are distinguished from those which are syntactically optional and amongst the latter, those which are existentially quantified are distinguished from those which are not.

### 2.1 Indirect Prepositional Arguments in Contact Predicates

Consider the following sentences:

- (1) O georgos fortose to ahiro sto karo.  
the farmer.N load.PAST.3S the hay.A onto-the wagon  
“The farmer loaded the hay on the wagon”.
- (2) O georgos fortose to karo me ahiro.  
the farmer.N load.PAST.3S the wagon.A with hay  
“The farmer loaded the wagon with hay”.
- (3) I diadilotes psekasan tin mpogia sto agalma.  
the demonstrators.N.PL spray.PAST.3PL the paint.A onto-the statue  
“The demonstrators sprayed the paint onto the statue”.
- (4) I diadilotes psekasan to agalma me mpogia.  
the demonstrators.N.PL spray.PAST.3PL the statue.A with paint  
“The demonstrators sprayed the statue with paint”.
- (5) The farmer loaded the hay on the wagon.
- (6) The farmer loaded the wagon with hay.
- (7) The demonstrators sprayed the paint onto the statue.

(8) The demonstrators sprayed the statue with paint.

(1)-(4) are examples of Modern Greek contact predicates which participate in the so-called locative alternation phenomena (see, among others, (Dowty, 1991), (Rappaport and Levin, 1988), (Levin and Rappaport Hovav, 1991)). Alternations in Modern Greek with the locative verbs *fortono* (load) and *psekazo* (spray) are of the general form  $NP_k V NP_i [P NP_j] \rightarrow NP_k V NP_j [P NP_i]$ , where the indices denote referential identity. The main features of such verbs in Modern Greek is that they are morphologically identical and that they involve two arguments: one denoting a *location* and one denoting the *locatum* (*karo* (wagon)/*agalma* (statue) and *ahiro* (hay)/*mpogia* (paint), respectively, in (1)-(4) above).

Two arguments, one of which denotes a *location* and the other the *locatum* (*wagon/statue* and *hay/paint*, respectively, in (5)-(8) above), are also supported by the English contact predicates *load* and *spray*.

(Levin, 1993) describes this class of predicates as follows:

[Locative alternation] is found with certain verbs that relate to putting substances on surfaces or things in containers, or to removing substances from surfaces or things from containers.

Much of the discussion in the literature has dealt with the so-called holistic interpretation of the English locative verbs *spray* and *load*.

In (5) all the available hay has been loaded onto the wagon no matter whether the wagon is full or not. In (6) the wagon is completely loaded. Likewise in (7) all the paint has been sprayed on the statue which is not necessarily covered. In (8) all the statue is covered. The aspect of all the sentences in (5)-(8) above, though, depends on the properties of the object rather than the properties of the oblique.

Not all locative verbs in English, though, alternate.

The verbs *fill* and *cover*, for instance, admit a *with-PP* indirect prepositional argument only (see also (Levin, 1993)):

- (9) Peter filled the tank (with water).
- (10) \*Peter filled water (into the tank).
- (11) Peter covered the garden (with a tarpaulin).
- (12) \*Peter covered a tarpaulin (over the garden).

On the other hand, the verb *pour*, for instance, appears only with a locative prepositional argument:

- (13) Peter poured water into the bowl.
- (14) \*Peter poured the bowl with water.

## 2.2 Indirect Prepositional Arguments in Removal Predicates

*Removal* predicates in Modern Greek and English also take *locatum* and *location* arguments and they are distinguished in the following groups:

(1) Predicates which imply a change of state of the *location* argument when this is realized as the direct object of the verb. These predicates may appear as tri-valent with alternative argument structures:

- (15) O Petros adiase tin dexameni (apo to nero).  
the Peter.N empty.PAST.3S the tank.A (of the water)  
“Peter emptied the tank (of water)”.
- (16) O Petros adiase to nero apo tin dexameni.  
the Peter.N empty.PAST.3S the water.A from the tank  
“Peter emptied the water from the tank”.
- (17) Peter emptied the tank (of water).
- (18) Peter emptied the water from the tank.

(2) Predicates which denote a contact with the *location*, as well as a change of location. These predicates may also specify the manner or the instrument related to the action of moving. For instance, the Modern Greek removal predicate *skupizo* (wipe) does not admit an indirect prepositional argument (*apo*-PP complement) when its *location* argument is realized as its direct internal argument (object; example (19)). In this case *skupizo* does *not* entail the existence of a *locatum* argument. For instance, the act of wiping a pan does not necessarily result in wiping something off it.

The corresponding predicates in English do not allow an inchoative interpretation (example (21)). This is an indication that they do not imply a change of state of the *location* argument. For instance, wiping the oil from a pan does not imply a definite change of the state of the pan. That means that the pan is not an *oil-less pan*.

- (19) \*O Petros skupise to tigani apo to ladi.  
the Peter.N wipe.PAST.3S the pan.A from the oil  
“\*Peter wiped the pan of the oil”.
- (20) O Petros skupise to ladi apo to tigani.  
the Peter.N wipe.PAST.3S the oil.A from the pan  
“Peter wiped the oil from the pan”.
- (21) \*The pan wiped of oil.

(22) \*Peter wiped the pan of the oil.

(23) Peter wiped the pan.

(24) Peter wiped the oil from the pan.

(3) However, the removal predicates *katharizo* in Modern Greek and *trim* in English are different than *skupizo* and *wipe*, respectively, in the sense that “trimming an object” necessarily means “trimming something off this object”:

(25) O Petros katharise to thamno apo ta xera kladia.  
the Peter.N trim.PAST.3S the bush.A of the dry branches  
“Peter trimmed the bush of the dry branches”.

(26) Peter trimmed the bush of the dry branches.

## 2.3 Indirect Prepositional Arguments in Impingement Predicates

A typical impingement verb in Modern Greek is *htipo*. Its English counterpart is *hit*. According to (Dowty, 1991), the verb *hit* (in English) does not imply any change of state for any of its arguments which may surface syntactically as direct internal arguments (objects). The same semantic entailments also hold for the Modern Greek verb *htipo*. *hit*, as well as *htipo* in Modern Greek, are assymetric predicates in that when the *location* argument is realized as the direct internal argument (object) of the predicate the *locatum* argument is the optional indirect prepositional argument, but when the *locatum* argument is realized as the direct internal argument all arguments are obligatory.

(27) O Petros htipise ton frahti.  
the Peter.N hit.PAST.3S the fence.A  
“Peter hit the fence”.

(28) O Petros htipise ton frahti me to xilo.  
the Peter.N hit.PAST.3S the fence.A with the stick  
“Peter hit the fence with the stick”.

(29) O Petros htipise to xilo sto frahti.  
the Peter.N hit.PAST.3S the stick.A onto-the fence  
“Peter hit the stick against the fence”.

(30) \*O Petros htipise to xilo.  
the Peter.N hit.PAST.3S the stick.A  
“\*Peter hit the stick”.

(31) Peter hit the fence.

(32) Peter hit the fence with the stick.

- (33) Peter hit the stick against the fence.
- (34) \*Peter hit the stick.

For verbs in the *htipo/hit* subclasses of Modern Greek and English, the “*me/with*” alternant (examples (28) and (32)) entails that one of the arguments (i.e., the *locatum*) is understood as the instrument (“means”) which is used by the actor in order to perform the action denoted by the verb. The “*stol/against*” alternant (see examples (29) and (33)), on the other hand, entails that the *locatum* undergoes directed motion; it is moved by the actor into contact with the location.

### 3. Previous Accounts in HPSG

(Markantonatou and Sadler, 1996) use underspecified verb entries in order to provide an HPSG analysis for verb alternations in English which affect specifically the choice of direct and indirect internal arguments.

In their analysis no lexical rules are implicated in relating the two different semantics they assume for the English locative verbs, which correspond to different syntactic argument structures. Instead, for their analysis they rely on the application of the rules of their linking component, the simultaneous satisfaction of different constraints and on type inference.

As an example of how their analysis works, let us take a closer look at their proposal for the English verb *load*, which has two alternative forms, each with an optional oblique which is existentially quantified when not syntactically realized:

- (35) John loaded the hay on the wagon.
- (36) John loaded the wagon on the hay.

(37) below is the semantic representation that (Markantonatou and Sadler, 1996) assume for the (active) English verb *load*.

They presuppose that

“...the [English] verb *load* has **only one** argument for which properties relevant to linking are expressed. This argument is the argument which will eventually surface as the subject. Otherwise, *load* requires a location and a locatum argument, but it does not define any entailments over these arguments which would enforce any particular linking” (Markantonatou and Sadler, 1996, pg. 52).

According to (Markantonatou and Sadler, 1996), it is this lack of further specifications which permits the location-object locatum-object alternation, and which reflects the fact that the two alternants of the verb *load* in English are somehow symmetric with respect to the optionality of oblique arguments. As far as existential quantification is concerned, they assume that arguments which appear in the lexical entry of *load* as first level or embedded (second level) semantic arguments are existentially quantified. *load*, according to them, also

has a value specified for the attribute SEM.CONS, which indicates that there is an entailment of contact between the ARG1 and the ARG3 of the predicate *load* (the location and the locatum). (Markantonatou and Sadler, 1996) underline that “the fact that this is the most general type of contact will in turn ensure that the predicate can surface with both *with*-PP and *on, in, etc*-PP”.

(37)	<i>spec</i>	
REL	<i>load</i>	
ARG1	$\boxed{1} \begin{bmatrix} \text{argtype} \\ \text{OTHER } \{ \text{location} \} \end{bmatrix}$	
ARG2	$\begin{bmatrix} \text{argtype} \\ \text{LINK } \text{causer\_ntc} \\ \text{OTHER } \{ \} \end{bmatrix}$	
ARG3	$\boxed{2} \begin{bmatrix} \text{argtype} \\ \text{OTHER } \{ \text{locatum} \} \end{bmatrix}$	
SEM.CONS.	$\begin{bmatrix} \text{contact} \\ \text{REL } \perp \\ \text{ARG1 } \boxed{1} \\ \text{ARG2 } \boxed{2} \end{bmatrix}$	

As far as linking of the arguments of the verb *load* is concerned, (Markantonatou and Sadler, 1996) assume that by means of the semantic representation that they propose in (37) two options are possible: “[Either] ARG2 is linked to subject as it has no other choice, and since it is a top level argument which is not also the argument of an embedded predicate, it must be linked. [Or] ARG1 and ARG3 are not specified for any LINK values and therefore they can each link either to the object of the verb or to the object of a predicate that maps an embedded relation.... [Finally] similar argumentation can be developed if one assumes that instead of linking the ARGs first, the system links SEM.CONS first” (Markantonatou and Sadler, 1996, pg. 52-53).

Finally, the fragment of the hierarchy of *semcons* in Figure (1.1) below shows how the alternation characterizing the locative verbs like *load* in English is accounted for in the theory proposed by (Markantonatou and Sadler, 1996), which we have presented briefly above.

#### 4. Indirect Prepositional Arguments: The Analysis

The robust account we present here for indirect prepositional arguments in Modern Greek and English does not follow the analysis of such arguments that

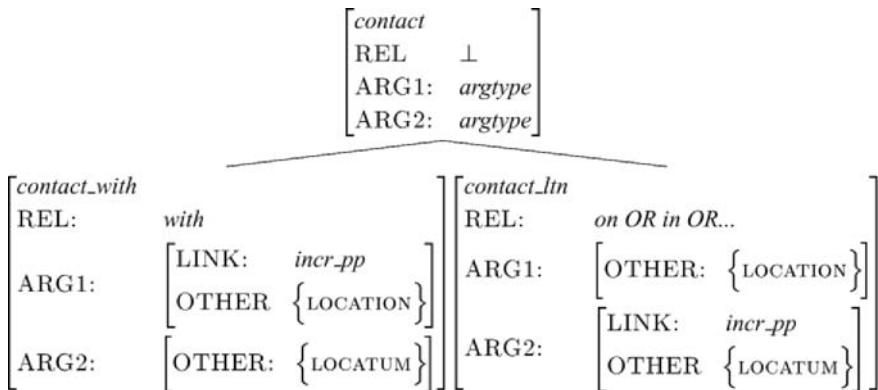


Figure 9.1. The hierarchy of SEMCONS that (Markantonatou and Sadler, 1996) propose for English locative verbs like load

(Markantonatou and Sadler, 1996) have proposed and whose main points we have briefly presented in Section 1.3 above. The reason is that we do not know how to deal and restrict the overgeneration that comes along with the radical underspecification (of verbal entries and/or of the arguments these support) that they assume.

The account proposed here for indirect prepositional arguments in Modern Greek and English (see examples in Section 1.2 above) is based on a minimal recursion approach to semantic representation and is formalized using the Minimal Recursion Semantics (MRS) framework of (Copestake et al., 1999) and (Copestake et al., 2001). In brief, Minimal Recursion Semantics is a framework for computational semantics, in which the meaning of expressions is represented as a flat bag of Elementary Predications (or EPs) encoded as values of a RELS attribute. The denotation of this bag is equivalent to the logical conjunction of its members. Scope relations between EPs are represented as explicit relations among EPs. Such scope relations can also be underspecified. The assumption of current MRS is that each lexical item (other than those with empty EP bags) has a single distinguished main EP, which is referred to as the *KEY EP*. All other EPs either share a label with the *KEY EP* or are equal to some scopal argument of the *KEY EP*. According to (Koenig and Davis, 2000), for situation-denoting EPs, which are also most interesting for our purposes here, the following generalizations hold: (i) EPs do not encode recursively embedded state-of-affairs (SOAs); (ii) EPs can have one, two, or three arguments; (iii) if an EP has three arguments, then one of them is a state-of-affairs, and another is an undergoer co-indexed with an argument of the embedded state-of-affairs. Finally, as far as direct arguments are concerned, these are predicted to link off the value of the *KEY* attribute.

Assuming that lexical items include more than one EPs in their semantic content, but lexically they select only one of these EPs as their KEY, we propose that the semantic properties of the arguments of the verb *fortono* (load) in example (2) of Section 1.2.1 are captured by the semantic type in (38).

(38) CONTENT value of *fortono* (load)

KEY	$\boxed{3} \left[ \begin{array}{l} \text{fortono-ch-of-st-rel} \\ \text{ACT } \boxed{1} (\text{o georgos}) \\ \text{UND } \boxed{2} (\text{to karo}) \end{array} \right]$
RELS	$\left\langle \boxed{3}, \left[ \begin{array}{l} \text{me-rel} \\ \text{ACT } \boxed{1} (\text{o georgos}) \\ \text{UND } \boxed{4} (\text{ahiro}) \\ \text{SOA } \boxed{3} \end{array} \right], \left[ \begin{array}{l} \text{fortono-ch-of-loc-rel} \\ \text{ACT } \boxed{1} (\text{o georgos}) \\ \text{FIG } \boxed{4} (\text{to ahiro}) \end{array} \right] \right\rangle$

(38) captures that the *me* (with) alternant of the Modern Greek locative verb *fortono* (load; example (2) in Section 1.2.1) denotes situations that must be both changes of state and changes of location.

The *sto* (onto) alternant of the Modern Greek locative verb *fortono* (load; example (1) of Section 1.2.1 above), on the other hand, denotes a single change of location.

We propose that the semantics of the *sto* (onto) alternant of the Modern Greek locative verb *fortono* includes only the last member of the RELS in (38) above.

This captures the CONTENT value of the *sto* (onto) alternant of the Modern Greek locative verb *fortono* (load) in example (1) of Section 1.2.1 as shown in (39) below.

The analysis presented above holds also for both alternants of the Modern Greek contact predicate *psekazo* (spray) (see examples (3)-(4) of Section 1.2.1 above and (40) and (41) below), as well as for both alternants of the English contact predicates *load* and *spray* (see examples (5)-(8) in Section 1.2.1 above).

(39) CONTENT value of *fortono* (load)

KEY	5	$\left[ \begin{array}{l} \text{fortono-ch-of-loc-rel} \\ \text{ACT } \boxed{1} (\text{o georgos}) \\ \text{FIG } \boxed{4} (\text{to ahiro}) \end{array} \right]$
RELS	$\langle \boxed{5} \rangle$	

(40) CONTENT value of *psekazo* (spray)

KEY	6	$\left[ \begin{array}{l} \text{psekazo-ch-of-st-rel} \\ \text{ACT } \boxed{1} (\text{i diadilotes}) \\ \text{UND } \boxed{2} (\text{to agalma}) \end{array} \right]$
RELS	$\langle \boxed{6}, \left[ \begin{array}{l} \text{me-rel} \\ \text{ACT } \boxed{1} (\text{i diadilotes}) \\ \text{UND } \boxed{4} (\text{mpogia}) \\ \text{SOA } \boxed{6} \end{array} \right], \left[ \begin{array}{l} \text{psekazo-ch-of-loc-rel} \\ \text{ACT } \boxed{1} (\text{i diadilotes}) \\ \text{FIG } \boxed{4} (\text{mpogia}) \end{array} \right] \rangle$	

(41) CONTENT value of *psekazo* (spray)

KEY	7	$\left[ \begin{array}{l} \text{psekazo-ch-of-loc-rel} \\ \text{ACT } \boxed{1} (\text{i diadilotes}) \\ \text{FIG } \boxed{4} (\text{mpogia}) \end{array} \right]$
RELS	$\langle \boxed{7} \rangle$	

As far as indirect prepositional arguments in removal predicates in Modern Greek and English are concerned, we propose that the semantic properties of the arguments of the verbs *skupizo* and *wipe*, for instance, which denote a change of location, when a *locatum* argument is realized as their direct internal argument (object; examples (20) and (24) of Section 1.2.2), can be captured by a type like the following (for Modern Greek the *rel* is *skupizo-ch-of-loc-rel* as shown in (42), while for English the *rel* would be *wipe-ch-of-loc-rel*):

(42) CONTENT value of *skupizo*

KEY	5	$\left[ \begin{array}{l} \text{skupizo-ch-of-loc-rel} \\ \text{ACT } \boxed{1} (\text{o Petros}) \\ \text{FIG } \boxed{4} (\text{to ladi}) \end{array} \right]$
RELS	$\langle \boxed{5} \rangle$	

*katharizo* (see example (25) in Section 1.2.2) is different than *skupizo*:

(43) CONTENT value of *katharizo*

KEY	③	[ <i>katharizo-ch-of-st-rel</i> ]	ACT ① ( <i>o Petros</i> )	UND ② ( <i>ton thamno</i> )	]
RELS	③,	[ <i>apo-rel</i> ]	ACT ①	[ <i>katharizo-ch-of-loc-rel</i> ]	]
			UND ④	ACT ①	
			SOA ③	FIG ④ ( <i>ta kladia</i> )	

That is, as (43) above captures, in Modern Greek trimming necessarily results in trimming something off something else; in the case of example (25) above trimming the bush results in trimming the dry branches off the bush. And this is what the semantic type in (43) captures. The semantic properties of the English verb *trim* in example 926) of Section (1.2.2) can also be captured by a type like the one in (43) adapted to English.

Finally, (45) below captures that the alternant of the Modern Greek impingement verb *htipo* (hit) whose indirect internal argument is headed by the preposition *me* (with; see example (28) in Section 1.2.3) entails that this argument, i.e., the *locatum*, is optional (see <sub>SOA</sub> (5) in (45)) and is understood as the instrument which is used by the actor in order to perform the action denoted by the verb. The same holds for the alternant of the English impingement verb *hit* whose indirect internal argument is headed by the preposition *with* (see example (32) in Section 1.2.3).

(44) below captures that the alternant of the English impingement verb *hit* whose indirect prepositional internal argument denotes the *location* (example (33) in Section 1.2.3) entails that the *locatum* (FIG(ure)) argument undergoes directed motion; it is moved by the actor into contact with the location. The same holds also for the alternant of the Modern Greek impingement verb *htipo* whose indirect prepositional internal argument denotes the *location* (see example (29) in Section 1.2.3).

(44) CONTENT value of *hit*<sup>1</sup>

KEY	7	$\left[ \begin{array}{l} \text{hit-directed-motion-to-contact-rel} \\ \text{ACT } \boxed{1} \text{ (Peter)} \\ \text{FIG } \boxed{4} \text{ (the stick)} \end{array} \right]$
RELS	$\left\langle \boxed{7}, \left[ \begin{array}{l} \text{against-rel} \\ \text{ACT } \boxed{1} \text{ (Peter)} \\ \text{UND } \boxed{8} \text{ (fence)} \\ \text{SOA } \boxed{7} \end{array} \right] \right\rangle$	

(45) CONTENT value of *htipo*

KEY	5	$\left[ \begin{array}{l} \text{htipo-rel} \\ \text{ACT } \boxed{1} \text{ (o Petros)} \\ \text{UND } \boxed{3} \text{ (ton frahti)} \end{array} \right]$
RELS	$\left\langle \boxed{5}, \left[ \begin{array}{l} \text{me-rel} \\ \text{ACT } \boxed{1} \text{ (o Petros)} \\ \text{UND } \boxed{4} \text{ (xilo)} \\ \text{SOA } \boxed{5} \end{array} \right], \left[ \begin{array}{l} \text{htipo-dmtc-rel} \\ \text{ACT } \boxed{1} \\ \text{FIG } \boxed{4} \end{array} \right] \right\rangle$	

## 5. Conclusion

In conclusion, we have shown that the theoretical framework of HPSG (Pollard and Sag, 1994) enriched with semantic representations in Minimal Recursion Semantics (MRS; Copestake et al., 1999; Copestake et al., 2001) constitutes the appropriate theoretical basis for a robust, linguistically-motivated account of indirect prepositional arguments, which provides the necessary formal generalizations for the analysis of such arguments in a multilingual context, since MRS structures are easily comparable across languages. To show this we have considered indirect prepositional arguments in contact, removal and impingement predicates in Modern Greek and English (Sections 1.2 and 1.4). Of course, as has already been shown in (Kordoni, 2003) and (Kordoni and Neu, 2003), the analysis presented in Sections 1.2 and 1.4 can be adapted and extended accordingly in order to account unproblematically for indirect prepositional arguments in German, as well.

As a final general comment we need to underline that the MRS-based analysis we have presented in Section 1.4 above allows for a linguistically motivated account of the syntactic properties of apparent semantic doublets, which avoids the processing load problems that are inseparable from (directional or even bi-directional (Flickinger, 1987)) lexical rule approaches to parsing indi-

rect prepositional arguments in particular and to development of (the lexicon of) large-scale deep computational grammars of natural language in general.

Consequently, (the lexicon of) large-scale computational grammars becomes more efficient, since it needs to depend on fewer or even no lexical rules at all, and thus less complicated for the grammar writer to maintain, as well as to develop further. Here we have focussed only on (some of) the theoretical assumptions upon which the achievement of such a goal can be based realistically.

## Acknowledgments

The comments of Julia Neu on earlier drafts have made this a better paper than it might otherwise have been.

## Notes

1. FIG(URE) denotes the moving entity (*locatum*); GRND (GROUND) denotes the contacted location (Davis, 2001).

## References

Baker, M. (1997), Thematic Roles and Syntactic Structures. In L. Haegeman (Ed.), *Elements of Grammar. Handbook of Generative Syntax*, pp. 73–137. Kluwer, Dordrecht.

Bender, E., D. Flickinger, and S. Oepen (2002), The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In J. Carroll, N. Oostdijk, and R. Sutcliffe (Eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics. Taipei, Taiwan*, pp. 8–14.

Callmeier, U. (2000), PET – a platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering* 6(1): Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation, 99–108.

Copestake, A. (2002), *Implementing Typed Feature Structure Grammars*. CSLI Lecture Notes, Number 110. Standord: CSLI Publications.

Copestake, A., D. Flickinger, I. A. Sag, and C. J. Pollard (1999), Minimal Recursion Semantics: An Introduction. Stanford University.

Copestake, A., A. Lascarides, and D. Flickinger (2001), An Algebra for Semantic Construction in Constraint-based Grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*, pp. 252–259. Toulouse, France.

Davis, A. (2001), *Linking by Types in the Hierarchical Lexicon*. Studies in Constraint-Based Lexicalism. Stanford: CSLI Publications.

Dowty, D. (1991), Thematic Proto-Roles and Argument Selection. *Language* 67, 547–619.

Flickinger, D. (1987), *Lexical Rules in the Hierarchical Lexicon*. Ph. D. thesis, Stanford University, California.

Jackendoff, R. (1990), *Semantic Structures*. Cambridge, Massachusetts: MIT Press.

Koenig, J.-P. and A. R. Davis (2000), The KEY to Lexical Semantics. Paper presented at the 7th International Conference on Head-Driven Phrase Structure Grammar, held on July 22-23, 2000 as part of the Berkeley Formal Grammar Conference.

Kordoni, V., (2003), The key role of semantics in the development of large-scale grammars of natural language. In *Proceedings of EACL'03, 10th Conference of the European Chapter of the Association for Computational Linguistics, Research Notes and Demos*, April 12-17, 2003, Budapest, Hungary, pp. 111–114.

Kordoni, V. and J. Neu (2003), Deep Grammar Development for Modern Greek. In *Proceedings of the Workshop on "Ideas and Strategies for Multilingual Grammar Development" taking place during the 15th European Summer School in Logic Language and Information (ESSLLI 2003)*, pages 65–72, Vienna, August 18-29, 2003.

Levin, B. and M. Rappaport Hovav (1991), Wiping the Slate Clean: A Lexical Semantic Exploration. In B. Levin and S. Pinker (Eds.), *Lexical and Conceptual Semantics*, pp. 123–152. Blackwell, Cambridge MA and Oxford UK.

Levin, B. and M. Rappaport Hovav (2001), What Alternates in the Dative Alternation? Ms., Colloquium Series, Department of Linguistics and Philosophy, MIT, Cambridge, MA, November 9, 2001.

Maling, J. (2001), Dative: The Heterogeneity of the Mapping Among Morphological Case, Grammatical Functions, and Thematic Roles. *Lingua* 111, 419–464.

Markantonatou, S. and L. Sadler (1996), Linking Indirect Arguments. *Essex Research Reports in Linguistics* 9, 24–63.

Mueller, S. and W. Kasper (2000), HPSG Analysis of German. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 238–253. Springer.

Oepen, S. and J. Carroll (2000), Performance Profiling for Parser Engineering. *Journal of Natural Language Engineering* 6(1): Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation, 81–97.

Pinker, S. (1989), *Learnability and Cognition: the acquisition of argument structure*. Cambridge, MA: MIT Press.

Pollard, C. and I. A. Sag (1994), *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Rappaport, M. and B. Levin (1988), What to do with  $\theta$ -roles. In W. Wilkins (Ed.), *Thematic Relations. Syntax and Semantics 21*, pp. 7–36. Academic Press Inc.

Williams, E. (1980), Predication. *Linguistic Inquiry 11*, 203–238.

# Chapter 10

## THE SYNTAX OF FRENCH *À* AND *DE*: AN HPSG ANALYSIS

Anne Abeillé<sup>1,5</sup>, Olivier Bonami<sup>1,4</sup>, Danièle Godard<sup>1,3,5</sup>, Jesse Tseng<sup>2,3</sup>

<sup>1</sup>*LLF UMR 7110*, <sup>2</sup>*Loria UMR 7503*, <sup>3</sup>*CNRS*, <sup>4</sup>*Univ. Paris IV*, <sup>5</sup>*Univ. Paris 7*

{abeille,dgodard} @linguist.jussieu.fr, olivier.bonami@paris4.sorbonne.fr, Jesse.Tseng@loria.fr

**Abstract** We present a descriptive overview of the uses of the French prepositional forms *à* and *de* and the properties of the constructions they appear in. The complexity of the data argues against a unitary syntactic and/or semantic treatment, but the empirical facts can nevertheless be organized in a systematic fashion. Concentrating primarily on *de*, we show that its uses can be grouped into two classes, one in which *de* patterns with ordinary prepositions, and one in which it appears to reflect the syntactic category of its sister element. We show that these observations can be accounted for in an HPSG analysis that distinguishes ordinary heads from ‘weak’ heads.

**Keywords:** French prepositions, weak heads, HPSG, extraction, coordination, portmanteau forms, pronominal clitics, grammatical marking

### 1. Introduction

The forms *à* and *de* have a large number of uses in French, with a complex array of syntactic properties. Since the facts for *à* and *de* are quite parallel, in this paper we will focus mostly on *de*, which has a wider range of functions. Where appropriate, we will point out details specific to *à*.

As the following examples illustrate, *de* can combine with an NP (46a). In certain cases the sequence [*de* LE N'] (where LE stands for all forms of the definite article) forms a so-called ‘partitive’ NP (46b). *De* can combine with an N' in a variety of contexts, with or without further combination with a quantifier to the left (47). It also combines with PPs (48), VPs (49), and APs and AdvPs (50).<sup>1</sup>

(46) a. Aller [de la gare] à l'hôtel. Un ami [de Marie].  
‘Go from the station to the hotel. A friend of Marie’s.’

- b. Demander [de la bière]. (de+LE ‘partitive’)  
‘Ask for beer.’

(47) a. Je n'ai pas lu [de journal]. (with negation)  
‘I didn't read any newspaper.’

- b. Il a lu beaucoup [de livres]. (with quantifier)
- c. Il a beaucoup lu [de livres]. (floating quantifier)  
‘He has read a lot of books.’
- d. Combien a-t-il lu [de livres] ? (extracted quantifier)  
‘How many books has he read?’

(48) Il surgit [de derrière l'église].  
‘He jumps out from behind the church.’

(49) Je me souviens [d'avoir lu ce poème].  
‘I remember having read this poem.’

(50) a. Quelqu'un de très fiable. J'ai encore trois jours [de libres].  
‘Someone very reliable. I have another three days free.’

- b. Quelque chose [de mieux], une page [de plus].  
‘Something better, one page more.’

In addition, in various cases, an idiosyncratic form appears where a combination of *de* and some other item would be expected; these cases are traditionally analyzed as post-syntactic reductions. (51a) illustrates the portmanteau forms found in place of *\*de+le* and *\*de+les*. As shown in (51b), the portmanteau form *des* can be reduced to *de* (or *d'*)<sup>2</sup> before a prenominal modifier. Finally, the single form *de* obligatorily replaces the ungrammatical combination of *de* followed by the ‘partitive’ article—i.e., *\*de+du/des/de la/de l'* (51c).<sup>3</sup>

(51) special realizations

- a. Acheter [du vin] / [des livres] (\*[de le vin] / \*[de les livres]).  
‘Buy wine/books.’
- b. Acheter [des/de beaux tableaux].  
‘Buy beautiful paintings.’
- c. Avoir besoin [d'aide] (\*[de de l'aide]),  
parler [de choses sérieuses] (\*[de des choses sérieuses]).  
‘Need help, talk about serious matters.’

The analysis of *de* (and to a lesser extent, of *à*) is a well-known puzzle of French grammar, and raises problems both from a semantic and a syntactic point of view. One question is whether *de* is always semantically empty (Blinkenberg, 1960), (Gougenheim, 1959), (Spang-Hanssen, 1963), (Cadiot, 1997)—see also the summary in (Kupferman, 1996)—and if not, whether

it can be treated as a polysemous element, with an abstract core meaning (Moignet, 1981), or whether there are several homophonous items. Syntactically, it is generally agreed that *de* does not belong to a single category (preposition); certain uses of *de* belong to some other category, or categories. However, there have been some attempts at a unitary analysis. (Milner, 1978, pp. 246–251) suggests that *de* (as well as *à*) could be a preposition in all of its uses, but not uniformly the head of a PP, thus dissociating the syntactic category from its habitual grammatical function. In recent terms, the preposition could be a ‘marker’ in some cases (Pollard and Sag, 1994); such an analysis is also compatible with the proposal of (Van Eynde, 2004) concerning some uses of Dutch prepositions. On the other hand, (Miller, 1992) suggests that *de* and *à* are phrasal affixes rather than prepositions—an analysis that sidesteps the syntactic category problem, but the problem of accounting for divergent properties of different uses of the same affix remains.

Three questions must be addressed. First, if one assumes a distinction between prepositional and non-prepositional uses of *de*, where is the dividing line between the two? For instance, when followed by an AP, should *de* be analyzed as a preposition (Azoulay-Vicente, 1985), or not (Huot, 1981)? When followed by an infinitival VP, is it always a preposition, always a complementizer, or one or the other depending on the environment (Huot, 1981)? Second, what is, or what are the categories of *de* when it is not a preposition, or does not head a PP? Different terms have been proposed for *de* in contexts where it does not seem to have the status of a normal preposition—“signe de liaison” (Blinkenberg, 1960), “cheville syntaxique” (Damourette and Pichon, 1911), “indice d’infinitif” (Gougenheim, 1959), “case marker” (Milner and Milner, 1972), (Vergnaud, 1974)—but no precise syntactic analysis has been offered. And finally, a question all too often neglected: How can we account for the common properties shared by both prepositional and non-prepositional uses of *de*?

The analysis we develop in this paper distinguishes two classes of uses of *de*, which we refer to as ‘oblique’ and ‘nonoblique’ uses. This partition is shown to be motivated by explicit syntactic criteria (and not correlated with semantic contentfulness). In oblique uses, represented by examples (46a), (48), and (50) above, *de* patterns with ordinary prepositions. On the other hand, in nonoblique uses, corresponding to (46b), (47), and (49), it does not behave like a normal preposition, and thus calls for a special analysis. For certain grammatical processes, however, we show that oblique and nonoblique uses pattern together, and in these cases a unified treatment is to be preferred.

The syntactic data justifying the distinction between the two types of uses are discussed in the following section. Section 3 offers an HPSG analysis of the empirical observations.

## 2. Syntactic properties

In this section we examine the syntactic behavior of the various constructions involving *de* illustrated in (46–50). We first show that a number of contrasting properties (possible syntactic function, extraction, wide scope over coordination) motivate a division between ‘oblique’ and ‘nonoblique’ uses of *de*. We then present two properties that cut across the oblique/nonoblique distinction (portmanteau forms, pronominal clitics) pointing to the existence of some unifying property common to all uses of *de*.

### 2.1 Syntactic functions

*De*-phrases can have a wide variety of syntactic functions, including complement and modifier of a number of different categories. Of particular importance is the fact that certain *de*-phrases can be subjects; this is the case for ‘partitive’ NPs and *de*-VPs (52).<sup>4</sup> Since PPs cannot be subjects in French (53), this observation casts doubt on any analysis that treats these *de*-phrases as PPs, and motivates our term ‘nonoblique’ for referring to this class of uses.<sup>5</sup>

- (52) a. [Des bijoux] ont été volés. ‘Jewels were stolen.’
- b. [De sortir un peu plus] te ferait du bien.  
‘Getting out a bit more would do you good.’
- (53) \*[Sous le lit] est un endroit idéal pour se cacher.  
‘Under the bed is an ideal place to hide.’

### 2.2 Extraction from *de*-phrases

The two types of *de*-phrases do not have the same properties with respect to extraction. Extraction out of nonoblique nominal phrases is possible (54a–c); notice that this is also what we observe for simple NPs not embedded in a PP (54d). By contrast, extraction is not possible out of oblique *de*+NPs or *de*+PPs (55a,b), just as with ordinary PPs (55c).<sup>6</sup>

- (54) a. Voici un auteur dont [des livres \_\_\_\_] sont en vente ici.  
‘Here’s an author some of whose books are on sale here.’
- b. Voici un auteur dont je n’ai pas lu [de livre \_\_\_\_].  
‘Here’s an author who I haven’t read any books by.’
- c. Voici l’auteur dont j’ai lu beaucoup [de livres \_\_\_\_].  
‘Here’s an author who I’ve read a lot of books by.’
- d. Voici un auteur dont j’aime [les livres \_\_\_\_] (mais pas les poésies).  
‘Here’s an author whose books I like (but not his poetry).’
- (55) a. \*Voilà le pays dont Paul revient [de la capitale \_\_\_\_].  
‘That’s the country that P.’s returning from the capital of.’

- b. \*Voilà l'homme dont Jean a surgi [de derrière la voiture \_\_\_\_].  
‘That’s the man whose car Jean jumped out from behind.’
- c. \*Voilà le livre dont j’ai déjeuné [avec l’auteur \_\_\_\_].  
‘That’s the book that I had lunch with the author of.’

We can apply the same test to other uses of *de*, with clear results. *De*-VPs pattern with nonoblique uses (56), while *de*-APs pattern with oblique ones (57).<sup>7</sup> Note that extraction out of *de*-AdvPs cannot be tested since the adverbs involved in this construction—recall the examples in (50b)—do not take (extractable) complements.

- (56) Voici le livre que je rêve [de traduire \_\_\_\_].  
‘Here’s the book that I dream of translating.’
- (57) a. Il n’y avait que Pierre [de convaincu de cette solution].  
‘There was only Pierre (who was) convinced by this solution.’
- b. \*C’est une solution dont il n’y avait que Pierre [de convaincu \_\_\_\_].  
‘That’s a solution that only Pierre was convinced by.’

## 2.3 Wide scope over coordination

Coordination provides another argument for distinguishing oblique and nonoblique uses. Nonoblique *de* never takes wide scope over a coordination of phrases, be they LE+N’ sequences in ‘partitive’ NPs, bare N’s, or VPs.

- (58) a. Pour ce gâteau, il faut de la farine et \*(de) la levure.  
‘For this cake, you need flour and baking powder.’
- b. Il y avait des pêches mûres et \*(des) tomates appétissantes.  
‘There were ripe peaches and appetizing tomatoes.’
- (59) Il y avait sur la table beaucoup de pain et \*(de) vin.  
‘There was a lot of bread and wine on the table.’
- (60) Je rêve de lire ce livre et \*(de) l’expliquer à mon fils.<sup>8</sup>  
‘I dream of reading this book and explaining it to my son.’

In contrast, oblique *de* can take wide scope over a coordination of NPs, PPs, APs, or AdvPs. It should be noted that judgments here are somewhat unclear, due to poorly understood semantic constraints and speaker variation.

- (61) J’ai besoin de [cette farine et cette levure] pour mon gâteau.  
‘I need this flour and this baking powder for my cake.’
- (62) Il revient de [chez Paul ou chez Marie].  
‘He’s coming back from Paul or Marie’s.’
- (63) a. quelqu’un de bon en maths et (de) fort en gym  
‘someone good at math and strong in P.E.’

- b. quelque chose de [plutôt bien ou plutôt mal]  
'something pretty good or pretty bad'

## 2.4 Portmanteau forms

One well-known property of *de* is its interaction with the definite article *le* (but not the elided form *l'*) and the plural *les* to yield the contracted or ‘portmanteau’ forms *du* and *des*—recall the examples in (51). The corresponding portmanteaux for *à* are *au* and *aux*. This phenomenon is completely regular for all instances where *à* and *de* appear in combination with *le* or *les*—irrespective of the oblique or nonoblique status of *de*.

On the other hand, *à* and *de* do not give rise to portmanteau forms when they combine with a VP, although the accusative pronominal clitics *le* and *les* are phonologically identical to the forms that trigger contraction in nominal contexts:

- (64) J'essaie [de les vendre] / \*[des vendre].  
'I am trying to sell them.'

This might be taken as evidence for distinguishing pre-verbal and pre-nominal *de*, but on the other hand the pronominal clitics *le* and *les* could simply be different from the articles *le* and *les*, in disallowing contraction.<sup>9</sup>

## 2.5 Pronominal clitics

It is well known that nominal phrases marked by *de* alternate with the pronominal clitic *en*. Interestingly, this same clitic is used regardless of the other properties of the phrase; in particular *en* is used both for oblique (65a–c) and nonoblique (65d–f) *de*-phrases.<sup>10</sup>

- (65) a. Je viens [de Londres] ⇒ J'en viens  
'I'm coming from London / from there.'
- b. Je me souviens [de ce poème] ⇒ Je m'en souviens  
'I remember that poem / it.'
- c. Je veux changer [d'hôtel] ⇒ Je veux en changer  
'I want to change (hotels).'
- d. Je n'ai pas [d'argent] ⇒ Je n'en ai pas  
'I don't have any (money).'
- e. J'ai trop [de travail] ⇒ J'en ai trop  
'I have too much (work).'
- f. Tu as acheté [de la bière] / [des livres] ⇒ Tu en as acheté  
'You bought some (beer/books).'

The clitic *en* is also available for some *de*-VPs, but not (for example) those that alternate with direct NP objects (Gross, 1975), (Huot, 1981):

(66) a. Je rêve [de venir demain] / [de cela] / \*cela  $\Rightarrow$  J'en rêve  
     'I dream of coming tomorrow / of that / of it.'

b. Venir demain, Paul en rêve.  
     'Coming tomorrow, Paul dreams of it.'

(67) a. Je promets [de venir] / \*[de cela] / cela  $\Rightarrow$  \*J'en promets  
     'I promise to come / that.'

b. Venir demain, Paul le promet.  
     'To come tomorrow, Paul promises (it).'

The VP data in (66) and (67) indicate that *de*-marked VP[*inf*] complements do not give rise directly to clitics. Instead, the main verb selects the form *le* or *en* (or *y* for à-marked complements) according to its NP or PP complementation frame (with further semantic restrictions). A VP[*inf*] can be anaphorically linked to this nominal or prepositional clitic, as in (66b) and (67b).

## 2.6 Interim conclusion

To sum up our observations so far, we have shown that uses of *de* are partitioned into two classes. When *de* precedes a PP, an AP/AdvP, or most NPs, it behaves like an ordinary preposition: The resulting phrase is an oblique complement or an adjunct, extraction is disallowed, and *de* can have wide scope over a coordination. On the other hand, when *de* precedes a VP, an N' (with the exception of examples like (65c), cf. fn. 10.4), or when it forms part of a so-called ‘partitive’ NP, it has properties that are unusual for prepositions: The resulting phrase can be a subject, extraction is possible, and *de* cannot take scope over a coordination. Finally, two properties are common to both classes: Oblique and nonoblique uses of *de* give rise to portmanteau forms and *en*-cliticization.

Before we turn to our analysis, two comments are in order. First, as stated in the introduction, the distribution of à is very similar to that of *de*, except for the fact that its range of uses is much more limited. À is always oblique in combination with NPs and PPs, and nonoblique only in combination with VP[*inf*]; the contrast is briefly illustrated below with data involving wide scope over coordination.

(68) a. J'ai parlé à Jean et (à) Marie.

    'I talked to Jean and (to) Marie.'

b. Jean a commencé à lire ce livre et \*(à) le traduire.

    'Jean has begun to read this book and translate it.'

Second, it is well known that *à* and *de* have uses where they are semantically contentful and uses where they are not. This semantic distinction does not coincide with the oblique/nonoblique division. This is illustrated in (69), where both examples are oblique uses. In combination with the copula (69a), *de* clearly expresses a semantic relation (here, ‘origin’). In (69b) however, *de* makes no semantic contribution: the semantic role of experiencer or stimulus is clearly a reflex of the lexical semantics of the noun.

(69) a. Paul est [de Paris]. ‘Paul is from Paris.’  
 b. la peur [des araignées]  
 ‘fear of spiders / the fear experienced by (the) spiders’

In many cases it is more difficult to establish whether or not *de* makes a semantic contribution. In particular we make no claim about the semantic contribution (or lack thereof) of nonoblique uses of *de*. But these examples show that semantic vacuousness and non-obliqueness must be treated as independent properties.

(70) Summary of empirical results

	‘oblique’ <i>à/de+NP/PP</i> <i>de+AP/AdvP</i>	‘nonoblique’ <i>à/de+VP</i> <i>de+N /NP</i>
has the distribution of	PP	VP/NP
extraction out of marked phrase	no	yes
wide scope over coordination	yes	no
portmanteau forms		yes
<i>en</i> -cliticization (of <i>de</i> -phrases)		yes
semantic contribution		sometimes

### 3. Proposed HPSG analysis

In this section we present an analysis of *de* (and *à*) that explicitly formalizes the difference between oblique and nonoblique uses, and at the same time provides a way to handle the properties they have in common. The analysis relies crucially on the novel concept of a ‘weak head’.

#### 3.1 Oblique uses: true prepositions

We treat *à* and *de* in their oblique uses as prepositions—i.e., as syntactic heads of category P, selecting a complement and projecting a PP. (71) is a description of the type *prep-word* subsuming all French prepositions. The HEAD value indicates the syntactic category (preposition), which propagates to all projections of lexical entries of this type. The MARKING attribute will be discussed in detail below in section 3.3. The empty SLASH set in (71) prevents extraction of and subextraction out of the preposition’s complement (if

any), because heads amalgamate the SLASH information of all their dependents (Bouma, Malouf, and Sag, 2001). In other words, French PPs are extraction islands. The COMPS list is left unspecified in (71) because French prepositions have quite diverse complementation frames, including intransitive uses: *Qu'est-ce que vous prendrez avec (cela)?* ‘What will you have with that?’; *Il y a de l'alcool dedans* ‘There's alcohol in it.’

$$(71) \quad \text{prep-word} \Rightarrow \begin{bmatrix} \text{HEAD} & \text{prep} \\ \text{MARKING} & \text{marked} \\ \text{SLASH} & \{ \} \end{bmatrix}$$

In our analysis, oblique *de* is of type *prep-word*. Thus it projects a PP from which nothing can be extracted. Moreover, oblique *de* is subject to the additional constraints in (72). It takes a COMPS-saturated complement, which is obligatory (*Qu'est-ce que vous faites dépendre de \*(cela)?* ‘What do you want to follow from that?’). Furthermore, this complement cannot be the projection of a verb (but can be nominal, prepositional, adjectival, or adverbial); consequently, *de*-marked infinitival VPs are exclusively nonoblique (as discussed in the next section).

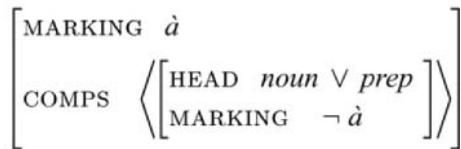
(72) oblique *de*: *prep-word* &

$$\begin{bmatrix} \text{MARKING} & \text{de} \\ \text{COMPS} & \left\langle \begin{bmatrix} \text{HEAD} & \neg \text{verb} \\ \text{MARKING} & \neg \text{de} \\ \text{COMPS} & \langle \rangle \end{bmatrix} \right\rangle \end{bmatrix}$$

Finally, (72) prohibits the complement of oblique *de* from bearing the MARKING value *de*. This blocks the ‘cacophonous’ repetition of *de* in examples like (51c) above (Gross, 1967). For instance, the *de*-marked ‘partitive’ phrase *de l'aide* cannot appear as the complement of the preposition *de*: *\*besoin de de l'aide*. To account for the so-called ‘haplography’ of *de* in the grammatical realization *besoin d'aide*, we assume a special lexical entry for prepositional *de* (or *d'*) selecting an N’ complement.

Oblique uses of à are also analyzed as prepositions (73). Prepositional à selects an obligatory nominal or PP complement, with a MARKING restriction to prevent repetition (*\*aller à à la station* ‘go to (at) the station’).<sup>11</sup>

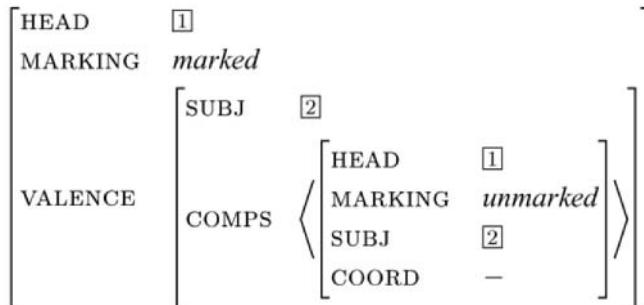
(73) oblique  $\grave{a}$ : prep-word &



### 3.2 Nonoblique uses: Weak heads

To handle the properties of nonoblique  $\grave{a}$  (appearing only in combination with infinitival VPs) and nonoblique *de* (appearing with VP[*inf*], LE+N', and bare N'), we appeal to the notion of ‘weak head’, which replaces the syntactic category of *marker* in classical HPSG (Tseng, 2002). A weak head is a lexical head that shares its syntactic category and other HEAD information with its complement. Recall that the HEAD value in HPSG encodes the part of speech (subtypes of *head*: *noun*, *verb*, *adj*, *prep*, etc.) and syntactic features appropriate for each part of speech (such as CASE, VFORM, PFORM, MOD). The sharing of HEAD values is indicated by the label 1 in the constraint in (74).<sup>12</sup> This accounts directly for the fact that certain properties of the non-head daughter remain visible on the phrase headed by nonoblique  $\grave{a}$  or *de*. For example, a control verb like *essayer* ‘try’ selects a complement headed by the weak head *de*, bearing the HEAD feature [VFORM *inf*]; if *de* were a true prepositional head in this construction, verb form information would be inaccessible for external selection.

(74) *weak-head*  $\Rightarrow$



The constraint in (74) further requires that the weak head inherit the subject list of its complement. This allows, for example, the subject of the verb in an  $\grave{a}$ - or *de*-marked VP[*inf*] to be controlled by the governing predicate: in other words, weak heads are subject raisers. Finally, the constraint [COORD −] on the complement prevents weak heads from taking wide scope over a coordinated structure.

It is crucial to note that the constraint in (74) says nothing about specifier valence. This allows the various forms of nonoblique *de* to constrain their specifier requirements in different ways. For some instances of the weak head *de*, we assume sharing of the SPR lists of the head and its complement, as indicated in (75a). Thus when *de* combines with an *N'*, it and the resulting *de*-marked *N'* remain SPR-unsaturated. Such unsaturated phrases are correctly predicted to have limited distribution; for example, they cannot appear in pre-verbal subject position (75b). We assume that the quantifier in structures of the form [Q *de N'*] is a specifier of *de*, giving rise to a fully saturated NP that can be a preverbal subject (75c).

(75) a. nonoblique *de*: weak-head &

HEAD	<i>noun</i> $\vee$ <i>verb</i>
MARKING	<i>de</i>
SPR	1
COMPS	$\langle [ \text{SPR } 1 ] \rangle$

b. \*Jean ne croit pas que [d'hommes] soient venus.  
 ‘Jean doesn't believe that any men came.’

c. [Beaucoup d'hommes] sont venus. ‘Many men came.’

For examples where *de-N'* does appear in post-verbal argument position, without an adjacent degree quantifier—recall examples (47a,c,d)—we propose an analysis relying on the idea of specifier extraction. The weak head *de* introduces an element in the SLASH set that must eventually be bound by a quantifier or by negation. See (Abeillé, Bonami, Godard, and Tseng, 2004) for a full presentation of our account of syntactic licensing of *de-N'* phrases.

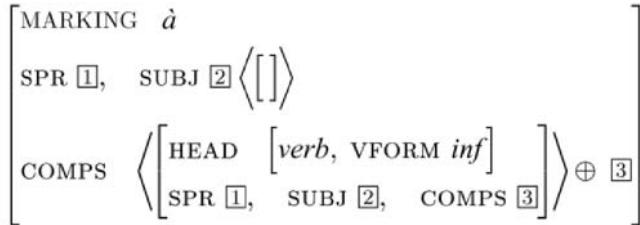
We analyze the so-called ‘partitive’ determiners *du*, *des*, *de la*, and *de l'* uniformly as synthetic forms (i.e., single lexical items).<sup>13</sup> Moreover, they are weak heads that select an *N'* complement (lacking a specifier), but they themselves do not require a specifier. This allows us to account for the NP-like distribution of ‘partitives’ while maintaining the generalization that we are dealing with *de*-marked phrases.

(76) *des, du, de la, de l'* : weak-head &  
 ‘partitive’

MARKING	<i>de</i>
SPR	$\langle \rangle$
COMPS	$\langle [ \text{HEAD } \text{ noun } \langle [ ] \rangle ] \rangle$

The lexical description of the weak head *à* is much more straightforward (77): it always selects a VP[*inf*] complement (from which it inherits all valence requirements) and introduces the MARKING value *à*.<sup>14</sup>

(77) nonobl. *à*: *weak-head* &



An alternative HPSG analysis for similar data is proposed by (Van Eynde, 2004), who distinguishes major and minor prepositions in Dutch. Like weak heads, minor prepositions introduce a specific MARKING value and do not contribute any part of speech information to the phrase. Unlike weak heads, they do not have syntactic head status, but are instead FUNCTOR daughters, like other non-head selectors (specifiers, modifiers). Finally, we note that ideas suggestive of the weak head approach can be found in the notion of “functional head” (Hulk, 1996) and in quantifier analysis of *de* presented by (Kupferman, 2004). Although their proposals differ from ours in significant ways (no account of the VP data, emphasis on the quantificational semantics of *de*), in their analyses *de* is also compatible with complements of various categories.

### 3.3 Grammatical marking

Up to now we have shown how a formal distinction between true prepositional heads and weak heads can account for the differences between oblique and nonoblique uses of *à* and *de*. But we also need to handle phenomena (in particular, cliticization) where prepositions and weak heads pattern together. To do this we rely on the MARKING specification.

The MARKING feature is familiar from previous work in HPSG; it is the feature that allows phrases containing an explicit marker (e.g., a complementizer) to be distinguished from unmarked phrases. We adopt the proposals of (Tseng, 2002), simplifying the Marking Theory of standard HPSG. This approach eliminates the syntactic category *marker* and the type *head-marker-phrase* of (Pollard and Sag, 1994). Markers are analyzed as weak heads that select an unmarked complement, while introducing a new MARKING value on the phrases they head. In this approach, the propagation of MARKING information is uniformly head driven; MARKING is not a HEAD feature, however, and so it is not shared between weak heads and their complements.

We assume that nouns and verbs are [MARKING *unmarked*] in the lexicon, and that each preposition introduces a specific *marked* value (see (72)

and (73)). Thus MARKING takes over the role played by PFORM in earlier HPSG, making it possible for a head to select a complement headed by a specific preposition. Two subtypes of the *marked* value *à* are needed, in order to account for the different cliticization possibilities:  $\dot{a}_{dat}$  corresponds to dative clitics such as *lui* (*faire confiance [à Paul]*  $\Rightarrow$  *lui faire confiance* ‘to trust Paul/him’), and  $\dot{a}_{loc}$  to the ‘locative’ clitic *y* (*penser [à Paul]*  $\Rightarrow$  *y penser* ‘to think about Paul/him’). This approach to marking allows for an account of preposition selection where the same preposition is selected on the basis of its semantics in some instances (Bonami, 1999), and on the basis of its MARKING value in others (Tseng, 2001).

The weak heads versions of *à* and *de* also introduce MARKING values, chosen from the same set of *marked* subtypes as their prepositional counterparts—recall (75a), (76), and (77). The MARKING attribute therefore provides a simple mechanism for handling phenomena in which corresponding weak heads and prepositions behave identically. In particular, we can easily account for the fact that all *de*-phrases—oblique and nonoblique—alternate with the same clitic *en* (cf. (65)). We analyze (*se souvenir de Marie* as a PP, (*beaucoup lu*) *de livres* as an N'[acc], (*boire*) *de la bière* as an NP[acc], and (*envie*) *de dormir* as a VP, but all of these phrases share the feature [MARKING *de*], and so they can all be replaced by the same clitic *en*. We know of no previous proposal capable of capturing this generalization, which involves phrases that are otherwise so dissimilar with respect to all other syntactic and semantic features.

## 4. Concluding remarks

We have presented an empirical overview of constructions involving the forms *à* and *de* in French, and we have offered a number of proposals for their analysis in the framework of HPSG. Our account depends crucially on the distinction between oblique uses of *à* and *de* (where they are analyzed as ordinary prepositional heads) and nonoblique uses (where they are analyzed as weak heads). At the same time, oblique and nonoblique uses of *à* and *de* can still pattern together thanks to a common inventory of MARKING values. This multi-faceted treatment accommodates most, if not all, uses of *à* and *de* in French and their properties with respect to a wide range of phenomena, including extraction, coordination, and cliticization.

## Notes

1. We mention in passing two constructions studied by (Milner, 1978) that we cannot deal with in this paper for lack of space:

- (i) Je préfère celui-là, [de manteau]. (right dislocation of N')  
'I prefer that one, that coat.'
- (ii) Il lui est déjà arrivé [de ces expériences]. (generalized 'partitive' NP, cf.(46b))  
'He has already had this kind of experience.'

It should also be noted that some constructions that might appear to belong in class (47) above are in fact members of (46a). These include certain [de N'] structures (*changer de nom* ‘change names’) and [quantifier de NP] constructions (*beaucoup de ces maisons* ‘many of these houses’).

2. All instances of *de* are systematically realized as *d'* before vowels. We will not go into the details of French vowel elision in this paper.

3. (Baronian, this volume) notes that this substitution or reduction strategy for avoiding two successive occurrences of *de* is not available in Quebec French, and speakers simply have to avoid such constructions.

4. Nonoblique *de-N'* combinations of the class represented by example (47) can also appear as subjects, but only post-verbally: *Combien sont venus [d'étudiants]?* ‘How many students came?’

5. Note the plural agreement on the verb in (52a). This contrasts clearly with the situation in English, where subject PPs are possible but do not trigger agreement (*Between the trees is/\*are a good place to park*).

6. The PP island constraint is weakened in some varieties of French.

7. Notice that extraction out of bare VPs is possible in French (i), while extraction from VPs introduced by other prepositional forms is subject to speaker variation (ii):

- (i) Voici les livres que j'aimerais [lire \_\_\_\_] / que je dois [traduire \_\_\_\_]  
‘Here are the books that I'd like to read / that I have to translate.’
- (ii) % Voici les livres que j'insiste [pour lire \_\_\_\_] / que je suis partie [sans lire \_\_\_\_]  
‘Here are the books that I insist on reading / that I left without reading.’

The AP examples in (57) require some further explanation: (iii) shows that the *de*-AP in (57a) is a predicative complement of the verb (and not an NP modifier), and (iv) shows that extraction out of adjectival complements of a verb is possible.

- (iii) Il n'y avait [de convaincu] que Pierre.  
‘There was no one convinced but Pierre.’
- (iv) C'est une solution dont je le croyais [convaincu \_\_\_\_].  
‘That's a solution that I thought he was convinced by.’

Thus the only factor blocking extraction in (57b) is the presence of *de*.

8. Note that a single *à* or *de* can mark a VP containing a coordination of lexical Vs:

- (i) Je rêve de [lire et expliquer] ce livre à mon fils.  
‘I dream of reading and explaining this book to my son.’

See (Abeillé and Godard, 1997).

9. This is impossible to test directly, since NPs cannot begin with a clitic and VPs cannot begin with an article. Unfortunately, for lack of space, we cannot go into the complex problem of the realization and distribution of portmanteau forms, which (contrary to traditional assumptions) is not a purely phonological phenomenon. See (Baronian, this volume) for a discussion of the problem in Quebec French and standard French.

10. There are well-known restrictions on adnominal *en*, depending on the function of the NP where it originates (Milner, 1978).

11. The complement of oblique *à* is not required to be COMPS-saturated (compare (72) and (73)). Heads of complex predicates in French (such as the copula) can and sometimes must inherit the unrealized complements of their complements (Abeillé and Godard, 2002), (Abeillé, Godard, Miller, and Sag, 1997). An inherited complement can be realized as a clitic on the main verb (i). This is generally blocked if the upstairs complement is a PP (ii), but some cases of complement inheritance are accepted, by some speakers (iii):

- (i) Il est [tout fier de son exploit] ⇒ Il en est tout fier  
‘He is so proud of his accomplishment / of it.’
- (ii) Il est [à la plage d'Arcachon] ⇒ \*Il en est à la plage.  
‘He is on the beach of Arcachon / of it.’
- (iii) Il est [à l'origine du canular] ⇒ % Il en est à l'origine.  
‘He is at the origin of the hoax / of it.’

12. Note that weak heads differ from functional heads in LFG or GB, for example. Although a weak head's category is underspecified in the lexicon, in any given syntactic context, it has a completely ordinary syntactic category (e.g. N or V). It is important to emphasize that when a weak head inherits a `HEAD` value of type *verb* or *noun*, it does not actually 'become' a verb or a noun (i.e., a lexical object of type *noun-word* or *verb-word*). Thus it is not surprising that it behaves very differently from a normal noun or verb with respect to complementation, inflection, etc.

13. Recall that the form *des* alternates with a reduced form *de* (51b); we will not formalize here the phonosyntactic conditions governing this alternation. Furthermore, our system would also permit an analytic treatment of *de la* and *de l'*, but we have seen no strong evidence in favor of this approach. Finally, note that for lack of space we omit the analysis of the portmanteau forms *du* and *des* in oblique (prepositional) uses. In oblique uses, the sequences *de l'* and *de la* are naturally analyzed as two-word combinations.

14. An example of complement inheritance by nonoblique à is in *tough* constructions like *facile à lire* 'easy to read'. Here, the unrealized direct object of *lire* is inherited by à and is therefore visible on the phrase à *lire*, where it can be selected by the adjective *facile* (Abeillé, Godard, Miller, and Sag, 1997).

## References

Abeillé, A., Bonami, O., Godard, D., and Tseng, J. (2004) The syntax of French *de-N'* phrases. In S. Müller (ed), *Proceedings of the 11th International Conference on HPSG*, pages 6–26. Stanford, CA: CSLI Publications.

Abeillé, A., and Godard, D. (1997) The syntax of French negative adverbs. In D. Forget, P. Hirschbühler, F. Martineau, and M.-L. Rivero (eds), *Negation and polarity: Syntax and semantics*, pages 1–27. Amsterdam: John Benjamins.

Abeillé, A., and Godard, D. (2002) The syntactic structure of French auxiliaries. *Language*, 78:404–452.

Abeillé, A., Godard, D., Miller, P., and Sag, I. A. (1997) French bounded dependencies. In S. Balari and L. Dini (eds), *Romance in HPSG*, pp. 3–56. Stanford, CA: CSLI Publications.

Azoulay-Vicente, A. (1985) *Les tours comportant l'expression de + adjectif*. Geneva: Droz.

Blinkenberg, A. (1960) *Le Problème de la transitivité en français moderne : essai syntactico-sémantique*. Copenhagen: Munksgaard.

Bouma, G., Malouf, R., and Sag, I. A. (2001) Satisfying constraints on extraction and adjunction. *Natural Language and Linguistic Theory*, 19:1–65.

Bonami, O. (1999) *Les constructions du verbe : le cas des groupes prépositionnels argumentaux*. PhD thesis, Université Paris 7.

Cadiot, P. (1997) *Les prépositions abstraites en français*. Paris: Armand Colin.

Damourette, J., and Pichon, E. (1911) *Des Mots à la pensée. Essai de grammaire de la langue française*. Paris: Editions d'Artrey.

Gougenheim, G. (1959) Y a-t-il des prépositions incolores en français ? *Le Français Moderne*, 27(1):1–25.

Gross, M. (1967) Sur une règle de cacophonie. *Langages*, 7:105–119.

Gross, M. (1975) *Méthodes en syntaxe*. Paris: Hermann.

Hulk, A. (1996) L’ “autre” *de* : une tête quantificationnelle ? In L. Kupferman (ed), pp. 44–59.

Huot, H. (1981) *Constructions infinitives du français : le subordonnant de*. Geneva: Droz.

Kupferman, L. (1996) Présentation. In L. Kupferman (ed), pp. 3–8.

Kupferman, L., ed. (1996) *Un bien grand mot : de. De la préposition au mode de quantification*. *Langue Française* 109.

Kupferman, L. (2004) *Domaines prépositionnels et domaines quantificationnels : Le mot de*. Paris: Duculot.

Miller, P. (1992) *Clitics and constituents in Phrase Structure Grammar*. New York: Garland.

Milner, J.-C. (1978) *De la syntaxe à l’interprétation*. Paris: Le Seuil.

Milner, J., and Milner, J.-C. (1972) La morphologie du groupe nominal en allemand. *DRLAV* 2. Université Paris 8.

Moignet, G. (1981). *Systématique de la langue française*. Paris: Klincksieck.

Pollard, C., and Sag, I. A. (1994) *Head-Driven Phrase Structure Grammar*. Stanford, CA: CSLI Publications. Distributed by University of Chicago Press.

Spang-Hanssen, E. (1963) *Les Prépositions incolores du français moderne*. Copenhagen: G. E. C. Gads Forlag.

Tseng, J. (2001) *The Selection and Representation of Prepositions*. PhD thesis, University of Edinburgh.

Tseng, J. (2002) Remarks on marking. In F. Van Eynde, L. Hellan, and D. Beermann (eds), *Proceedings of the 8th International Conference on HPSG*, pp. 267–283. Stanford, CA: CSLI Publications.

Van Eynde, F. (2004) Minor Adpositions in Dutch. *Journal of Comparative Germanic Linguistics*, 7:1–58, 2004.

Vergnaud, J.-R. (1974). *French relative clauses*. PhD thesis, MIT.

# Chapter 11

## IN SEARCH OF A SYSTEMATIC TREATMENT OF DETERMINERLESS PPs

Timothy Baldwin,<sup>1</sup> John Beavers,<sup>2</sup> Leonoor van der Beek,<sup>3</sup> Francis Bond,<sup>4</sup> Dan Flickinger<sup>2</sup> and Ivan A. Sag<sup>2</sup>

<sup>1</sup>*University of Melbourne and NICTA Victoria Laboratory*, <sup>2</sup>*Stanford University*,

<sup>3</sup>*Groningen University*, <sup>4</sup>*NTT Communication Science Laboratories*

tim@cs.mu.oz.au, {jbeavers,danf,sag} @csli.stanford.edu, vdbeek@let.rug.nl,  
bond@cslab.kecl.ntt.co.jp

**Abstract** This paper examines determinerless prepositional phrases in English and Dutch from a theoretical perspective. We classify attested P + N combinations across a number of analytic dimensions, arguing that the observed cases fall into at least three distinct classes. We then survey four different analytic methods that can predict the behaviour of the differing classes and examine various remaining difficult cases that may remain as challenges.

**Keywords:** determinerless PP, multiword expression, selection, noun countability

### 1. Introduction

There is a growing appreciation of multiword expressions (MWEs) as an obstacle to automated language understanding (Sag, 2002a), (Calzolari et al., 2002). In this paper, we highlight some of the peculiarities of MWEs, focusing on determinerless prepositional phrases (PPs). We then outline an analysis that can be used to systematically handle the phenomenon.

**Determinerless PPs** (henceforth PP–Ds) are defined to be made up of a preposition (P) and a singular noun (N<sub>Sing</sub>) without a determiner (Quirk et al., 1985) (Huddleston et al., 2002, as in Table 11.1, organised roughly by semantic type (cf. (Stvan, 1998). In the case that the noun is countable (e.g. *by bus*, *in mind*), a syntactically-marked structure results as the noun in itself does not constitute a saturated NP. This poses a problem for both parsing and generation unless we have some explicit treatment of this unexpected grammaticality. Orthogonally, PP–Ds can occur with idiosyncratic semantics (e.g. *at bay* and

*in kind*) which a system must have prior knowledge of to be able to analyse correctly.

PP–Ds exist in most languages with articles, and the same semantic types appear in a variety of languages: English, Albanian, Tagalog and German to name just a few (Himmelmann, 1998). Articles are generally used less frequently and less consistently in adposition phrases than in other syntactic environments. However, articles are regularly omitted in expressions of similar semantic types across languages: institution/location (*at school*), metaphor/abstract (*at large*), temporal (*in winter*), means/manner (*by car*). In this paper we will principally be concerned with English and Dutch data. Although the broad analysis is valid for other languages, the details will of course vary between languages (e.g. see Abeillé et al., (this volume) for an analysis of determinerless usages of *à* and *de* in French), and even across dialects of English (Chander, 1998).

Despite their regularities, PP–Ds tend to receive a simple ‘words with spaces’ treatment in lexical resources. **COMLEX**, for example, lists a total of 762 PP–Ds, in the form of a set of prepositions a given countable noun can occur with in a PP–D construction (Grishman, 1998). As **COMLEX** was developed as an exclusively syntactic resource, only syntactically-marked PP–Ds feature in the lexicon, and coverage tends to be patchy (e.g. in **COMLEX** 3.0, *tricycle* is listed as occurring in *via/by tricycle*, *motorbike* in only *by motorbike*, and *bicycle* has no annotated PP–D usages). **WordNet** (Fellbaum, 1998) is more ad hoc in its treatment of PP–Ds, listing around 80 PP–Ds in the adjective section and 330 in the adverb section. Predictably, the PP–Ds that are described in **WordNet** tend both to have predicative usages and to be semantically marked. The lexicon for the Japanese-to-English machine translation system **ALT-J/E** lists several classes of nouns that interact with prepositions and affect article usage, such as institutions and meals Bond:2001. However, the list is far from complete, and the classes are not explicitly linked to semantic classes.

To get a preliminary sense for the extent of the problem posed by PP–Ds and the relative success of **COMLEX** and **WordNet** at listing them, we carried out a semi-automated analysis of PP–D occurrences in the written component (80m words) of the British National Corpus (BNC, (Burnard, 2000)), using the method described in (Baldwin, 2003a)<sup>1</sup>. Focusing on the prepositions *as*, *at*, *by*, *in* and *on*, we first manually inspected all extracted PP–Ds which occurred at least 20 times in the corpus, and removed syntactically and semantically unmarked PPs (e.g. *at midnight*). These post-corrected sets were used to estimate the type and token coverage of **COMLEX** and **WordNet** over PP–D types in the BNC. Based on the relative error rates in each of these sets, we estimated the type and token frequencies of PP–Ds occurring at least 5 times in the BNC. The final results are presented in Table 11.2.

Institution	Media	Metaphor	Temporal	Means/Manner
<i>at school</i>	<i>lon film</i>	<i>on ice</i>	<i>at breakfast</i>	<i>by car</i>
<i>in church</i>	<i>on TV</i>	<i>at large</i>	<i>at lunch</i>	<i>by train</i>
<i>in gaol</i>	<i>to video</i>	<i>at hand</i>	<i>on break</i>	<i>by hammer</i>
<i>on campus</i>	<i>off screen</i>	<i>at leave</i>	<i>by night</i>	<i>by computer</i>
<i>at temple</i>	<i>in radio</i>	<i>at liberty</i>	<i>by day</i>	<i>via radio</i>
...	...	...	...	...

Table 11.1. Examples of English PP–Ds, classified according to the system of (Stvan, 1998)

	FREQUENCY $\geq 20$						FREQUENCY $\geq 5$			
	BNC		In COMLEX		In WordNet		Types		Tokens (%)	
	Types	Tokens	Types	Tokens	Types	Tokens	Types	Tokens	Types	Tokens (%)
<i>as</i>	41	7,292	0%	0%	0%	0%	484	12,686	(0.02%)	
<i>at</i>	54	18,948	15%	17%	22%	59%	289	28,580	(0.04%)	
<i>by</i>	71	8,327	35%	48%	1%	1%	1,023	15,493	(0.02%)	
<i>in</i>	237	113,235	29%	45%	9%	14%	1,918	113,582	(0.13%)	
<i>on</i>	99	25,097	26%	44%	7%	9%	964	28,204	(0.04%)	

Table 11.2. Coverage and corpus occurrence of English PP–Ds

The coverage figures for **COMLEX** and **WordNet** vary according to the preposition, but **COMLEX** tends to have a token coverage of around 30% and **WordNet** a token coverage of around 15%, underlining the inadequacies of the two lexical resources with respect to PP–Ds. Turning next to the type and token frequency estimations, it becomes apparent that PP–Ds are a significant phenomenon in the BNC (accounting for over 0.2% of all tokens<sup>2</sup>). In summary, PP–Ds are surprisingly common in corpus data, and are treated inconsistently in lexical resources.

The remainder of the paper is structured as follows. We describe the syntax and semantics of PP–Ds in sections 2 and 3 respectively. In section 4 we sketch our analysis. Our results are summarised in section 4.5.

## 2. The Syntax of Determinerless PPs

The syntax of PP–Ds is not uniform. The constructions differ in their level of syntactic markedness, productivity and modifiability. On the one extreme, we have (typically Latinate) MWEs that are historically P + N combinations (*ex cathedra*, *ad hominem*, *ad nauseum*, etc.) but which, despite the erudition of certain speakers, are still best analysed as fixed expressions (Sag, 2002a). These constructions are non-productive and non-modifiable. On the other extreme are fully productive and modifiable combinations of P + complement, where lexical selections<sup>3</sup> interact with a general head-complement construc-

tion to build standard PPs with compositional semantics (*per recruited student that finishes the project*). Much of English lies in between these two extremes.

We classify PP–Ds primarily in terms of their syntactic markedness, dependent largely on the nature of the prepositions and the uses of the nouns outside of these PPs. Syntactically unmarked PP–Ds are those where the  $N_{Sing}$  can occur without a determiner outside of the PP (i.e. the  $N_{Sing}$  is uncountable). For example, some prepositions select for an argument that is unbounded (uncountable or plural countable), e.g. *out of generosity* in English and *uit vrijevigheid* “out of generosity” in Dutch. The determinerless nature of these PPs is not surprising and since these PPs are not marked syntactically (and often not semantically either as we’ll discuss in the next section) they do not pose a significant problem for a (computational) grammar. A more interesting group is institutions (the social/geographic spaces in (Stvan, 1998)), which appear to be semi-productive. Some prepositions like *in* can combine with a range of these nouns (*in church, in school, in court, in goal*), although other members of the same semantic class are not necessarily possible (\**in library*, although context often improves these readings). However, this contrast mirrors the contrast between *school is over* and \**library is over*: the nouns that can appear in this type of PP–D can also appear without a determiner outside of PPs, and in this way these PP–Ds are not syntactically marked.

On the other hand, there are prepositions that specifically require their argument to be both determinerless and countable, causing the PP to be syntactically marked. An example is the preposition *per* in both English and Dutch. Most prepositions do not specify the countability of their argument, so that the PP–Ds are sometimes syntactically marked (with a countable noun) and sometimes unmarked (with an uncountable noun). For example, means/manner *by* as in *by car, by computer*, takes a wide, productive class of normally countable nouns that almost never occur without determiners. These are syntactically marked in the sense that the noun otherwise would require a determiner. But the same preposition combines with an uncountable noun in the syntactically unmarked PP *by public transportation*.

Another factor relevant to syntactic markedness is modifiability, and here most PP–Ds lie in the middle of the spectrum (Ross, 1995). Except for the fixed expressions mentioned above, most PP–Ds are modifiable to some extent. At the two extremes of modifiability are PP–Ds that allow no modification at all (*of course, in \*children’s/\*mental/\*small hospital*<sup>4</sup> and Dutch *in principe* “in principle”) and PP–Ds that obligatorily require modification (*at great/public/considerable expense, for good/safe measure and op vreemde/Nederlandse bodem* “on foreign/Dutch soil”, but not \**at expense, for measure* or \**op bodem* “on soil”). Between these two extremes, some PP–Ds only allow idiosyncratic modification (*at long/\*great/\*short last*), while others allow modification more freely (*at great/considerable/tedious/epic length*).

Overall, though, modification is seldom unrestricted (in which case it tends to occur with fully productive constructions, e.g. *per recruited student that finishes the project* (from above)), and on this criterion virtually all PP–Ds are somewhat marked.

We can get some sense of the distribution of PP–Ds across the spectrum of relative modifiability by analysing the probabilistic predictability of modification patterns of different PP–D types. This is achieved via a process of cross-validation, whereby we partition up the BNC data into 10 contiguous segments of equal size, and compare the distribution of modifiers for a given PP–D in each of the 10 segments as compared to the remaining 9 segments. We determine the normalised distribution of modifiers in each case and calculate the Kullback Leibler (KL) divergence between the two distributions to determine their relative fit, averaging over the 10 iterations of cross-validation to attain a single divergence ( $D$ ) value. Where the two distributions are identical, i.e. the exact same modifiers occur with the same relative occurrence, the KL divergence is 0, and failing this the magnitude of the divergence reflects the relative mismatch of the two distributions. In practice, there is a high correlation between the relative scope of modification and the KL divergence value as relative freedom of modification gives rise to greater variance in both the range of modifiers observed in a given partition and the relative frequency of each. By correlation, therefore, PP–Ds with low KL divergence have restricted modifiability, and tend to occur unmodified the bulk of the time. In addition to analysing KL divergence relative to other instances of the same PP–D ( $D(PP \parallel PP)$ ), we calculate the divergence over NPs not selected for by prepositions ( $D(PP \parallel NP)$ ). This provides some insight into the relative markedness of modification relative to non-PP occurrences of the same noun. That is, we would expect to see relative low divergence for productive PP–Ds due to their greater compositionality, and relatively high divergence for PP–Ds with marked syntax and/or semantics.

In Table 11.3, we present a random sample of 20 PP–Ds occurring with frequency 100 or greater in the BNC, in increasing order of  $D(PP \parallel PP)$ . Items higher in the list can be seen to resist modification, which in the case of *horse-back*, e.g., is consistent with its behaviour outside of PP–Ds, whereas with *contrast*, the lack of modification appears particular to PP–Ds. At the end of the list, we see that with *on analysis*, there is greater variability in modification within the PP–D data than relative to non-PP usages. The relative increase in the value of  $D(PP \parallel PP)$ , is slow, indicating that even for PP–Ds with scope for modifier variation, actual variation tends to be slight.

For PP–Ds that allow modification, there can be additional constraints on the word class of the modifier. Some PP–Ds allow only noun-noun compounds, as with *at eye/street level* but not *\*at higher level*, while others allow only adjective modifiers, as with *in sharp/pointed/rich contrast* but not *\*in*

Prepositional Phrase	Divergence	
	$D(PP \parallel PP)$	$D(PP \parallel NP)$
<i>on horseback</i>	0.00	0.04
<i>before dawn</i>	0.00	0.16
<i>in reverse</i>	0.00	0.51
<i>by contrast</i>	0.00	0.71
<i>to hospital</i>	0.02	0.32
<i>into bed</i>	0.02	0.56
<i>up front</i>	0.03	0.26
<i>by marriage</i>	0.05	0.29
<i>on trial</i>	0.07	0.21
<i>on record</i>	0.10	0.76
<i>in readiness</i>	0.11	0.50
<i>in diameter</i>	0.14	0.54
<i>in school</i>	0.18	0.26
<i>on loan</i>	0.18	0.71
<i>in isolation</i>	0.19	0.83
<i>in disgust</i>	0.22	0.34
<i>in depth</i>	0.27	0.50
<i>in tone</i>	0.87	1.08
<i>by decree</i>	1.62	2.07
<i>on analysis</i>	4.29	2.81

Table 11.3. A random sample of 20 PP–Ds occurring  $\geq 100$  times in the BNC

*color contrast*. The two dimensions of choice of modifier (noun, adjective, or either) and presence of the modifier (obligatory, impossible, or optional), combine to present seven logically possible subclasses of PP–Ds (since the subclass that disallows modifiers is indifferent to the dimension of modifier choice), as shown in Table 11.4. Each of these logically possible subclasses is instantiated in the BNC data. Other languages may have different constraints on modification. Some Dutch prepositions allow morphological but not syntactic modification, but select for a bare noun at the same time. Here, the prepositional object can only be modified via morphological rules, by forming a complex N (*op zeilkamp* “at sailing camp”, *op ponykamp* “at pony camp” and *op schoolkamp* “at school camp”, but not *\*op sportief kamp* “at sporty camp”).

	Obligatory	Optional	Impossible
Noun	<i>at *(eye) level</i>	<i>on (summer) vacation</i>	
Adjective	<i>at *(long) range</i>	<i>in (sharp) contrast</i>	<i>on (*very) top</i>
Either	<i>at *(company) expense</i> <i>at *(considerable) expense</i>	<i>in (family) court</i> <i>in (open) court</i>	

Table 11.4. Variation in modification of determinerless PPs

Despite this rich spectrum of syntactically distinct PP–Ds, there are still some constructions that don't seem to fit in. In the first place there are some prepositional constructions consisting of two prepositions with determinerless arguments: *from X to Y*, *X by X*, e.g. *from mother to child*, *room by room*. Secondly, features of determinerless constructions may be distributed over both conjuncts of a coordination where only one fulfils the selectional requirements of the preposition. For example, *in* does not readily occur with the noun *brush* in a PP–D, and yet the coordination *in brush and ink* is perfectly acceptable (noting that *in ink* is also a grammatical PP–D). Finally, there is a class of coordinated PP–Ds in Dutch where neither one of the coordinated nouns can occur independently in a determinerless PP (e.g. *over mens en wereld* “about human being and world”, *van stadion en hotel* “of stadium and hotel”).

### 3. The Semantics of Determinerless PPs

Turning to the semantics of PP–Ds, (Stvan, 1998) focused primarily on four natural semantic classes of nouns and a relatively small set of prepositions (mostly locatives like *at* and *on*), classifying them by possible implicatures (or enrichments of content) and contrasts with full NPs. However, looking at a broader set of data shows considerable systematicity along many other semantic dimensions, and in this section we will highlight some of these relevant categories and outline a general classification of PP–Ds based on semantic markedness. As noted above, all PP–Ds show a certain degree of markedness in the form of metaphorical (*on ice* in the non-literal sense), institutionalised (*at school*), and generic uses (*by car*), which in many (but not all) cases is different from the basic simplex semantics of these nouns. Relative to this, however, they seem to follow a cline of markedness dependent on both lexical semantics and the overall compositionality of the PP, with certain natural semantic classes often clustering together.

Among the least marked semantic classes of PP–Ds are those formed with institutionalised nouns such as *in town*, *at school*, *at church*, a sizeable subset of Stvan's social/geographic spaces, which in the previous section were identified as the least syntactically marked since they occur both in and out of PP–Ds without determiners. Corresponding to this distributional property, not surprisingly, are similar semantic effects. In PP–Ds, these show a variety of special semantics including what Stvan refers to as activity and familiarity implicatures. Activity implicatures (or enrichments of content) occur when the PP seems to be referring to an activity associated with the institution, rather than a specific place (e.g. *in gaol* “while being a prisoner” and *in school* “while attending school”, which can even be true of someone not located at a school, as opposed to *at a gaol/school* which is a simple locative). Familiarity arises from uses that seem to refer to specific entities familiar to a participant in the

discourse (e.g. *John is in town* “John is in (my/his) town”, as opposed to *John is in the/a town* which again is a simple locative).<sup>5</sup> However, most nouns in this institutionalised class have corresponding  $N_{Sing}$  non-PP uses that induce the same semantic effects, as in (78) (note that (78c) is particular to American English dialects where *school* can be synonymous with *university*):

- (78) a. *While at school*[=attending school], I learned the value of an education. (Complement of preposition)
- b. *School*[=attending school] drains the best years of your life. (Subject)
- c. *Many students can't afford school*[=to attend school] in the States. (Object)

In (78) each use of *school* can induce the same reading, in this case the activity [enrichment], and likewise for other uses, like familiarity [enrichment] (e.g. *work wore him out* where *work* can be replaced by *his work*, as well as *working*).<sup>6</sup> Given the persistence of this kind of specialised semantics, their universally determinerless nature, and the large size and semi-productivity of this noun class, the semantics of these PP–Ds is unsurprising and thus relatively unmarked, being entirely predictable from the N. The fact that institutional nouns can occur without determiners in these environments is, however, a peculiarity of English; related Germanic languages such as German and Swedish require the definite article here (Himmelmann, 1998). Dutch examples of institutional nouns that can occur in determinerless environments are *school* “school” and *kantoor* “office”. These examples show activity and familiarity implicatures similar to the English examples, but are less modifiable and less numerous. Norwegian has the intriguing property that PP–Ds tend to occur only in institutionalised contexts, e.g. the determinerless *i hengekøye* “in hammock” is grammatical only in combination with a verb such as *sove* “sleep” (Borthen, 2003).

Other nominal classes show varying degrees of semantic markedness, such as Stvan’s class of media expressions, e.g. *in print*, *on film*, *on video*, involving media-related nouns. Here, too, we see similar nominal semantics in and out of PPs:

- (79) a. *The Manchurian Candidate* is my favourite film.  
    [sense=content] [form=countable]
- b. *I'd rather watch it on film than rent the video.*  
    [sense=material] [form=uncountable]
- c. *I would always rather watch a film than a video.*  
    [sense=media form] [form=countable]

(Stvan, 1998)

In (79), *film* shows similar readings (specifically broadcast/media type, material, and content type) in a variety of positions, again showing a low degree of semantic markedness. However, unlike the institutional class, these uses rarely occur without determiners outside of PPs (although sometimes this is possible, e.g. *TV rots your brain* [sense=content]), indicating some degree of syntactic markedness. Another of Stvan's classes is "temporal interruptions", where the noun identifies a specific break in a particular routine, subdividing into two classes: shorter breaks marked by *at* (e.g. *at lunch*) and longer, more open-ended breaks with *on* (e.g. *on leave*). The nouns associated with short breaks occur frequently in similar uses outside of PP–Ds (e.g. *lunch starts at noon*), indicating less semantic markedness, whereas longer breaks involve nouns that rarely do (e.g. ??*vacation lasts longer each year*,<sup>7</sup> \**we want more holiday in our work year*), indicating more semantic markedness.

On the other end of the markedness scale is a class of non-compositional and relatively metaphorical PP–Ds, including *at hand* and *on ice*, largely corresponding to what Stvan labels "untethered metaphors", i.e. expressions formed by nouns that define states and generally have no referential properties. However, despite their non-compositionality, not all of these PPs are semantically unpredictable. In particular many adverbial and adjectival PP–Ds have synonymous, morphologically related adverb or adjective pairs, e.g. *lastly/at last, willfully/at will, effectively/in effect* and *handy/on/at hand, edgy/on edge*. While still idiosyncratic (e.g. *edgy/on edge* "nervy/exitable" is not entirely predictable from *edge*) the semantic relationship between these morphologically derived and analytic noun-centred forms is striking, showing some systematicity if not predictability.

Similarly, although prepositions do not cluster into fine-grained semantic classes like nouns, they show various semantic properties relevant to their distributions within PP–Ds. A significant number of spatial prepositions (e.g. *at, to, on*, etc.) occur in PP–Ds, in both temporal and stative uses, although this is hardly surprising since cross-linguistically spatial prepositions frequently grammaticalise into temporal and stative/metaphorical uses independent of PP–D constructions (correspondingly to a low degree of markedness) (see e.g. (Haspelmath, 1997)). However, there are further semantic dimensions within these broader semantic classes. For example, a variety of interesting patterns are seen in antonymous pairs of prepositions. With locative prepositions, several antonymous pairs show stark differences in their distribution, e.g. *on/off, in/out, at/away (from), near/far (from)*, etc. In our corpora, the inclusive or positive prepositions (e.g. *in, on*) were among the highest frequency heads while the negative pairs were generally much rarer (there were surprisingly few corpus examples involving *off, out* and *away (from)*, although these certainly do exist, e.g. *off base, away from town*). Interestingly, antonymous pairs for which neither preposition had an inclusive/positive reading tended to show

up infrequently, e.g. the relative infrequency of PP–Ds headed by *down/up*, *before/after*. Other anonymous pairs showed further interesting relationships. In our corpora, the relative frequency of *without* with uncountable nouns in generic readings (e.g. *without success*, *without fear*, *without help*) was roughly double that of *with*. Therefore it appears that cross-cutting semantic features such as inclusiveness/exclusiveness and negative polarity also play a role in the semantic regularity of PP–Ds. Synonymy, on the other hand, does not appear to be a relevant factor in determining grammaticality of PP–Ds. For example, *by* as in *by law*, where *by statute* is grammatical but not ??*according to law* and ??*according to statute*. This further highlights the generally lexicalised nature of PP–Ds. Crosslinguistically, primary adpositions (short monomorphemic adpositions with grammatical meanings) are more likely to be involved in PP–Ds than secondary adpositions (longer or complex adpositions with concrete meanings) (Himmelmann, 1998).

Finally, idiosyncratic prepositions sometimes form classes of PP–Ds all of their own. One of the most regular semantic classes is means/man-ner *by*, most of whose members are vehicular (e.g. *by car*, *by train*) although not always (e.g. *by hand*, *by post*, *by telephone*). In general these resist referential uses and familiarity enrichments, although they do allow generic and activity readings:<sup>8</sup>

(80) *I travelled to San Francisco by car. They're/It's a great way to travel/#It rattled a lot.*

Such PP–Ds tend to be nonreferential and more semantically marked than the institution class since most of these nouns rarely occur with the means/manner semantics in subject/object position (although it is possible, e.g. *car costs less than train for trips to the city*). On the other hand, this class shows a high degree of internal systematicity, particularly in excluding related readings with determiners (e.g. \**by a/the car*) and some amount of productivity (e.g. *I arrived yesterday by carpet* in a context of having a flying carpet – see section 4). These are just a few of the myriad levels of (semi-)regularity in the PP–D system. Although previous work has focused primarily on systematicity in relation to natural semantic class of the  $N_{Sing}$  and the small set of possible interpretations, it appears there is a wider set of generalisations, taking into account basic semantic features of the prepositions and broader lexical classes inside and outside of PP–Ds.

## 4. Analysis

As noted in the introduction, the coverage of existing resources is unsystematic and generally limited to more or less fixed preposition-noun combinations. We will introduce three more analyses to complement this: occurrence with defective noun phrases, selection for idiosyncratic noun phrases and selection for nominal phrases ( $\bar{N}$ s) by the preposition. Each of the four kinds of anal-

yses is well suited for a large class of PP–Ds. The analyses are given in the framework of Head-driven Phrase Structure Grammar, and have been tested by implementing them in the English Resource Grammar (Flickinger, 2000), although with only with a few examples of each kind.

In all three kinds of syntactic analysis, the familiar HPSG head-complement construction will license all the PP–Ds in question. But the differing lexical specifications will modulate the relevant distributions appropriately. For any given PP–D, there should always be evidence (modification, productivity) to tell if it is to be lexically listed or treated syntactically. If it is treated syntactically, then there should be further evidence showing whether the prepositional object is a freely combining NP, a modified nominal phrase with idiosyncratic restrictions on the presense or type of modifier or a specially selected, unsaturated nominal phrase ( $\bar{N}$ ) (the determinerless NP in non-prepositional contexts, restricted choice of P).

## 4.1 Lexical Listing

Lexical listing is the obvious approach for the syntactically and semantically marked class (e.g. *at large*, *on track*). For expressions such as these, it is entirely sufficient to simply list the P + N combinations in the lexicon, since the combination is non-productive and largely non-modifiable. In addition, the semantics is non-compositional and uniquely associated with a particular PP–D. Lexical listing is a simple approach that accurately reflects the inflexibility of these PPs.

For the other types of PP–Ds, lexical listing is more problematic. First, modification of the nominal within the PP can be possible (e.g. *as former president*, *at considerable length*). Simple lexical listing cannot handle this. Second, the syntactically marked class, e.g. *by car*, *by train*, *by taxi*, is productive, which also makes a simple listing in the lexicon impossible. Moreover, the semantically unmarked constructions have compositional semantics. Hence any attempt to treat the preposition and noun as a multiword lexical unit would fail to express this compositionality. Finally, some of the PP–Ds (or rather the nominals within them) select for an optional prepositional complement (e.g. *in front of the children*). This selection is also hard to capture via simple lexical listing.<sup>9</sup> Within a syntactic approach, one might consider positing a general rule:  $NP \rightarrow \bar{N}$ . However, such a rule would massively overgenerate, as any noun would be allowed to occur sans determiner in any context. Even if the rule were restricted to PP contexts, it would overgenerate, as not all prepositions and not all nouns allow the determinerless combination. Therefore, it would appear that a more fine-grained treatment is needed.

## 4.2 Prepositions that occur with Defective NPs

Some PP–Ds can be analysed as simple syntactic combinations of a preposition and an NP complement. The NP itself is defective and has no determiner. The key motivation for such an analysis, as noted in section 3 above, is the fact that these noun phrases appear without a determiner in other (semantically appropriate) syntactic contexts, e.g. as subjects and objects. For example, *church*, *school*, etc. are countable nouns that refer to (sets of) churches, schools, etc. But these give rise to the determinerless noun phrases *church*, *school*, etc. that refer to the relevant church and school activities: for example *School is over*, cited in section 2 above. Our account of these PP–Ds requires no new apparatus: since the determinerless noun phrases exist independently as subject and object NPs, it follows that they should also appear as prepositional objects in a standard head-complement construction. The semantics seems equally straightforward, in that the semantic composition of *in school* acquires the interpretation “in the appropriate school-related activity” in just the same way that *likes school* acquires its “likes the appropriate school-related activity” interpretation, as discussed in section 3. This analysis also predicts that the determinerless NP in question will not be restricted to a single preposition. Though certain P + N combinations may give rise to semantic incompatibility, the general prediction made by this analysis seems right for this class of expression, given that *in/at/after/before/during school* are all well-formed and easily interpretable.

## 4.3 Prepositions that select idiosyncratic NPs

Next, we present an analysis for the more idiomatic PP–Ds where the nouns can take only a restricted set of modifiers. In this case the idiosyncratically modified nouns also construct defective noun phrases, but they are constrained to only appear as complements of prepositions, as with *at eye level* or *at considerable expense*.<sup>10</sup>

The syntactic analysis employs three unary rules similar to the bare-NP rule used for constructing full determinerless NPs from ordinary mass or plural nominal phrases. For each of these three additional rules, the daughter is constrained to be headed by a particular subclass of nouns, idiosyncratically marked in the lexicon for the property of being modifiable by a noun, an adjective, or neither. Two of the three rules require that the daughter be a nominal phrase containing a (pre-head) modifier, while the third rule constrains the daughter to be unmodified. On this account, a phrase like *at eye level* is thus analyzed as a head-complement structure combining the ordinary preposition *at* with the determinerless NP *eye level*, where this NP is constructed via a unary rule which constrains the daughter to be lexically headed by a noun which permits nominal modification, and moreover this daughter must indeed contain a

modifier. The lexical entry for this idiosyncratic *level* is distinct from the entry for the ordinary count noun *level*, and is constrained so that (1) phrases that it projects will only appear as complements of prepositions, (2) its specifier (determiner) will never be expressed, (3) it must combine with a (pre-head) modifier before it can combine with the preposition, and (4) it can only appear with a nominal modifier, not an adjective. Of course, this analysis only ensures that the syntactic constraints are correctly imposed on these subclasses of PP–Ds containing modified nominals. We will still require additional semantic collocational constraints analogous to those for semi-productive idioms (cf. (Riehemann, 2001)), in order to reflect the collocational restrictions on which specific prepositions combine with which of these modified nouns, and which modifiers are possible.

#### 4.4 Prepositions that select $\bar{N}$

The approaches just sketched will not extend to the productive constructions discussed earlier (e.g. *by car*; *as president*) in which a particular preposition (or preposition class) selects for an exclusively countable noun that cannot project a determinerless NP in other syntactic contexts:

(81) a. *They arrived by train/plane/bus/pogo stick/hydro-foil* ...  
 b. <sup>\*</sup>*I really like train/plane/bus/pogo stick/hydrofoil*  
 c. <sup>\*</sup>*Train/plane/bus/pogo stick/hydrofoil could save us money.*

When there is no evidence that a PP–D contains an NP-projecting uncountable noun, then it makes sense instead to posit a lexical entry or lexical type of preposition constrained as in (82):

(82) 
$$\text{SYN} \left[ \text{CAT} \left[ \begin{array}{ll} \text{HEAD} & \text{prep} \\ \text{VAL} & \left[ \text{COMPS} \left[ \left[ \text{SPR} \langle \text{Det} \rangle \right] \right] \right] \end{array} \right] \right]$$

Prepositions of this type select a complement whose specifier is of type *Det*. As only nouns have specifiers of type *Det*, and NPs have an empty specifier, the complement is constrained to be an  $\bar{N}$ . By positing an entry of this sort for (one sense of) the preposition *by*, we can account for its special ability to combine with determinerless (unsaturated) nominal phrases that denote means/instruments but wouldn't normally occur in this interpretation. Crucially, in all such cases, the determinerless nominal is restricted to the preposition *by*, as predicted:

(83) <sup>\*</sup>*They arrived with/in/to train/plane/bus/hydrofoil/pogo stick* ...

These productive PP–Ds seem further restricted to particular semantic domains, e.g. *on* + MEDIUM or *by* + MEANS/INSTRUMENT. These restrictions could be the result of selection for specific semantic classes of nouns by the preposition or they could alternatively be interpretations entirely contributed by the preposition on top of the nominal semantics. The Dutch construction *in* + PIECE OF CLOTHING is ungrammatical with anything that is not established as clothing and thus seems to suggest the former. However, examples like *From the train station to Hogwarts is 15 minutes by broom* suggest that the preposition supplies the interpretation, although it is a matter of descriptive granularity and/or domain-specificity as to whether the noun enables a matrix transportation interpretation or not.<sup>11</sup>

## 4.5 Summary of Analysis

Finally, although we have suggested that there are three distinct kinds of analysis, there are a number of cases that present challenges to this simple picture of the world of PP–Ds. For instance, there are many different PP–Ds with the English nouns *sea* and *hand* or the Dutch nouns *zee* “sea” and *huis* “house”. These PPs are semantically unmarked (the meaning is fully compositional) but syntactically marked (the nouns do not occur without a determiner outside of PPs). These are distinct from the *by car* type in that the determinerless P + N combination is not restricted to a particular preposition (e.g. *at sea*, *to sea*, *from sea to ...*, *%by sea*, *\*in sea*, *\*over sea*, ...). Perhaps these are idioms, whose common properties must be relegated to linguistic history; or perhaps there is some fine-grained semantic analysis that will account for the restricted distribution in synchronic terms. The work of (Soehn, 2003) provides a third alternative: an analysis in terms of selectional restrictions imposed by the noun. Our hope is that no such stipulations are required within an adequate grammar: in each such case there is some factor or factors to be discovered that interacts with the pristine picture of PP–Ds that we have sketched here.

## 5. Conclusion

We have presented PP–Ds as a commonly occurring, highly varied form of multiword expression, and documented their idiosyncratic syntax and semantics. Depending on the type of PP–D, one of four analyses was proposed: simple lexical listing, occurrence of the preposition for independently existing determinerless NPs, selection for idiosyncratic determinerless NPs or selection for nominal phrases (Ns). The analyses we have outlined cover a wide area, but do have yet to be reconciled with the full range of idiosyncratic restrictions on P + N combination that have been observed in the literature.

We have implemented these analyses in a computational grammar. The next step in our research is to extract determinerless PPs from corpora in volume

and analyse each for such properties as modifiability and referentiality. Using this as a guide, we can determine the robustness of the proposed analyses over open data and build up a rich inventory of lexicalised PP–Ds to supplement existing resources.

## Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Ann Copestake and the anonymous reviewers for their valuable input on this research.

## Notes

1. Both countable and uncountable nouns were included in this data.
2. Here, the percentages are calculated relative to the total token count in the BNC, not just tokens of frequency  $\geq 5$ .
3. Prepositions typically (but not always) select for an NP complement.
4. *In hospital*, and the indicated judgements for modifiability, are particular to British English.
5. This enrichment of content, however, seems to be somewhat intertwined with the ‘activity implicature’, since you can have this anaphoric reference even in activity senses, as in *his hair went grey in gaol*, which could mean *his hair went grey while serving time in his gaol* thus showing both enrichments. In other cases this is necessarily the case, as in *they had a bad day at work[=working at their workplace]*. In this regard the data is somewhat murky.
6. This goes against Stvan, who argues that such nouns in subject position do not show familiarity, although as noted in fn. (11.5) the data in general isn’t so clear.
7. Acceptable in some American dialects
8. PPs headed by *by* (and *via*) are not the only means/manner PPs, e.g. *on foot*, however we assume that cases such as this, which are non-productive and idiosyncratic, should be lexicalised.
9. An alternative approach to these transitive PPs is to analyse them as complex prepositions (prepositions with spaces). According to this analysis, *on top* is similar to *inside*, except that the former selects for a PP[*of*] and the latter for a complement that is either an NP or a PP[*of*].
10. Note that we adopt a somewhat unconventional treatment of noun–noun compounds such as *eye level*, in treating the first noun as a modifier of the second.
11. Such an analysis could also be extended to cases of *from X to Y* and *like X like Y*, e.g. *from town to town* or *like father like son* by assuming *from/like* takes two complements, an N and a particular PP, providing the appropriate semantic relationship between them.

## References

Baldwin, Timothy, Beavers, John, van der Beek, Leonoor, Bond, Francis, Flickinger, Dan, and Sag, Ivan A. (2003). In search of a systematic treatment of determinerless PPs. In *Proc. of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France.

Bond, Francis (2001). *Determiners and Number in English, contrasted with Japanese, as exemplified in Machine Translation*. PhD thesis, University of Queensland, Brisbane, Australia.

Borthen, Kaja (2003). *Norwegian Bare Singulars*. PhD thesis, Norwegian University of Science and Technology.

Burnard, Lou (2000). *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Calzolari, Nicoletta, Fillmore, Charles, Grishman, Ralph, Ide, Nancy, Lenci, Alessandro, MacLeod, Catherine, and Zampolli, Antonio (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–40, Las Palmas, Canary Islands.

Chander, Ishwar (1998). *Automated Postediting of Documents*. PhD thesis, University of Southern California, Marina del Rey, CA.

Fellbaum, Christiane, editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Flickinger, Dan, Oepen, Stephan, Uszkoreit, Hans, and Tsujii, Jun’ichi (2000). Journal of natural language engineering (special issue on efficient processing with hpsg).

Grishman, Ralph, Macleod, Catherine, and Myers, Adam (1998). *COMLEX Syntax Reference Manual*. Proteus Project, NYU.

Haspelmath, Martin (1997). *From Space to Time in The World’s Languages*. Lincom Europa, Munich, Germany.

Himmelmann, Nikolaus P. (1998). Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology*, 2:315–353.

Huddleston, Rodney and Pullum, Geoffrey K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, and Svartvik, Jan (1985). *A Comprehensive Grammar of the English Language*. Longman, London, UK.

Riehemann, Susanne (2001). *A Constructional Approach to Idioms and Word Formation*. PhD thesis, Stanford, USA.

Ross, Háj (1995). Defective noun phrases. In *Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, pages 398–440.

Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, and Flickinger, Dan (2002). Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Soehn, Jan-Philipp and Sailer, Manfred (2003). At first blush on tenterhooks. About selectional restrictions imposed by nonheads. In Jäger, Gerhard,

Monachesi, Paola, Penn, Gerald, and Winter, Shuly, editors, *Proceedings of Formal Grammar 2003*, pages 149–161.

Stvan, Laurel Smith (1998). *The Semantics and Pragmatics of Bare Singular Noun Phrases*. PhD thesis, Northwestern University.

## Chapter 12

# COMBINATORIAL ASPECTS OF COLLOCATIONAL PREPOSITIONAL PHRASES

Beata Trawiński

*University of Tübingen, Collaborative Research Centre 441*

*Nauklerstraße 35, 72074 Tübingen, Germany*

[trawinski@sfs.uni-tuebingen.de](mailto:trawinski@sfs.uni-tuebingen.de)

Manfred Sailer

*University of Göttingen, Department of English Studies*

*Käte-Hamburger-Weg 3, 37073 Göttingen, Germany*

[manfred.sailer@phil.uni-goettingen.de](mailto:manfred.sailer@phil.uni-goettingen.de)

Jan-Philipp Soehn

*University of Tübingen, Collaborative Research Centre 441*

*Nauklerstraße 35, 72074 Tübingen, Germany*

[soehn@sfs.uni-tuebingen.de](mailto:soehn@sfs.uni-tuebingen.de)

**Abstract** In this paper we will discuss semantic aspects of collocational prepositional phrases (CPPs) consisting of  $P_1N_1P_2NP$  sequences. Based on the syntactic analysis in (Trawiński, 2003), which assumes prepositions heading  $P_1N_1P_2NP$  combinations to be able to raise and syntactically realize complements of their arguments, we will investigate whether the semantic representations of these expressions can be derived compositionally. We will discuss German CPPs with respect to two criteria of internal semantic regularity taken from (Sailer, 2003), and we will observe that the expressions in question are not uniform with regard to their semantic properties. While the logical form of some of them can be computed by means of ordinary meaning assignment and a set of standard derivational operations, others require additional handling methods. However, there are approaches available within the HPSG paradigm which are able to account for these data. Here we will briefly present the external selection approach of (Soehn, 2003) and the phrasal lexical entries approach of (Sailer, 2003), and we will demonstrate how they interact with the syntactic approach of (Trawiński, 2003).

**Keywords:** bound word, collocation, complex preposition, compositional semantics, HPSG, phrasal lexical entry, raising

Among *collocational prepositional phrases* (CPPs), sequences consisting of a preposition, a noun, a second preposition, and an NP ( $P_1N_1P_2NP$ ) occur particularly frequently in many languages.<sup>1</sup> These combinations are collocational in the sense of exhibiting a high degree of lexical fixedness. CPPs are commonly considered to be unpredictable with regard to standard grammar regularities. However, (Trawiński, 2003) has shown that the syntax of German CPPs can be described within HPSG (Pollard and Sag, 1994) using the well established mechanism of raising. Based on this syntactic approach, we will describe the semantic aspects of German CPPs. We will distinguish CPPs of different semantic regularity and combine independently motivated accounts to capture these expressions.

## 1. Syntactic Aspects

### 1.1 Some Empirical Observations

We consider the following word combinations to be  $P_1N_1P_2$  expressions.

(1) an Hand von (at hand of, 'by means of'), in Verbindung mit (in connection with, 'in connection with'), unter Aufsicht von (under survey of, 'under the supervision of') ...<sup>2</sup>

At first glance, the interdependence between the particular elements of these expressions seems to defy standard constraints on the PP structure of German; on examining PPs involving  $P_1N_1P_2$  sequences such as *in Verbindung mit* ('in connection with') in the contexts exemplified in (2), we can observe many differences compared to traditional PPs.

(2) In Verbindung mit diesem Problem will ich bemerken, dass ...  
in connection with this problem want I note that

'In connection with this problem, I want to point out that ...'

First of all, the noun *Verbindung* ('connection') cannot combine with a determiner, a quantifier, a possessive pronoun or a prenominal genitive (3a). Secondly, it cannot be modified (3b). Finally, the PP *mit diesem Problem* ('with this problem') cannot be omitted (cf. 3c).

(3a) \* in einer/ der/ seiner/ Peters Verbindung mit diesem Problem  
in a/ the/ his/ Peter's connection with this problem

(3b) in \*enger/ \*unerwarteter [Verbindung mit diesem Problem] \*von dieser Woche/ \*die uns betrifft, will ich ...  
in close/ unexpected [connection with this problem] from this week/ which us concerns want I

(3c) in Verbindung will ich ...  
 in connection want I

Based on these observations, it is often assumed that the string *in Verbindung mit* ('in connection with') is a complex lexical sign (cf. the structure in (4) provided for those PPs by (Fries, 1988)).

(4)  $[_P[_P[_P \text{ in}] [_N \text{ Verbindung}] [_P \text{ mit}]] [_NP \text{ diesem Problem}]]$   
 The preposition heading the entire phrase is a projection of three lexical categories which together form a complex lexical category, in this case a preposition *in Verbindung mit* ('in connection with'). This complex preposition then selects an NP forming a prepositional phrase.

The main problem with this analysis consists in the assumption that the preposition *mit* ('with') belongs to the complex preposition and cannot form a constituent with the NP *diesem Problem* ('this problem'). However, there are several data demonstrating the opposite.

Firstly, the combinations  $P_2 NP$  where  $P_2$  is realized by *von* ('of') can be replaced by the genitive; this replacement of *von* ('of') adheres to the restrictions on distribution of postnominal genitives and *von*-PPs in German (5a). Secondly, the sequences in question can be substituted by *wo-/da-* expressions as in (5b), which are usually considered as proforms for PPs. These observations imply that the  $P_2 NP$  sequences form a constituent.

(5a) an Hand von zwei Beispielen/ zweier Beispiele  
 by means of two examples/ two examples<sub>GEN</sub>  
 'by means of two examples'

(5b) in Verbindung womit/ damit  
 in connection WO\_with/DA\_with  
 'in connection with what/with it'

Taking all previous observations into consideration, one can conclude that within a  $P_1 N_1 P_2 NP$  expression the  $P_2 NP$  is lexically selected by  $N_1$ , but realized as a syntactic sister of a  $P_1 N_1$  complex.

## 1.2 Raising Analysis

Based on the above generalization, (Trawiński, 2003) provides an analysis for these expressions using the raising mechanism.<sup>3</sup> We will outline here the HPSG formalization of this analysis.

To avoid redundancies in the lexicon, only one lexical entry for *in* ('in') will be specified (cf. Figure 12.1), bearing underspecified information about its argument's degree of saturation. The syntactic selection properties of *in* ('in') are licensed by a constraint on the mapping of the elements of the ARG-ST list to the valence lists (cf. Figure 12.2). In order to enable prepositions to subcategorize nouns with an unsaturated complement, and then also to select the complements of those nouns, the list of complements which are syntactically

selected by a preposition is specified as a concatenation of its own ARG-ST list and the comps list of its argument (cf. 2  $\otimes$  1).

<i>word</i>
PHON $\langle$ in $\rangle$
ARG-ST $\langle$ [LOC   CAT   HEAD <i>noun</i> ] $\rangle$
SYNS   LOC   CAT   HEAD <i>prep</i>

Figure 12.1. The relevant part of the lexical entry of the preposition *in* ('in')

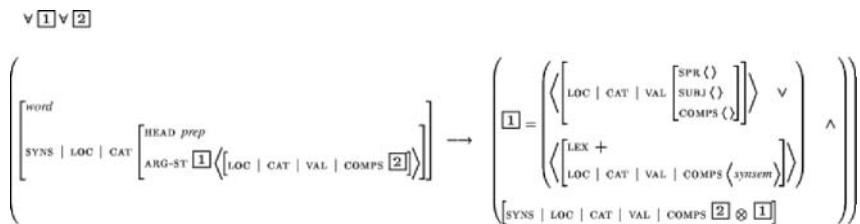


Figure 12.2. ARG-ST Mapping Lexical Principle for Prepositions

It should be mentioned that the raising of more than one nominal complement results in ungrammatical constructions such as the following:

(6) in Verbindung \*[der Regierung] mit diesem Problem ...  
in connection the government<sub>GEN</sub> with this problem

To avoid this problem the ARG-ST value of a preposition is restricted to be either a list with one saturated element, or a list containing one element with a singleton COMPS list (cf. 1). Additionally, the LEX value of the second disjunct is specified as +. This marks objects which have not realized any of their complements. This restriction rules out the selection of relational nouns which have already realized one of their complements (cf. 7).

(7) \* in [Verbindung der Regierung] [mit diesem Problem] ...  
in connection the government<sub>GEN</sub> with this problem

The structure in Figure 12.3 exemplifies the interaction of our assumptions regarding the licensing of a PP headed by a raising preposition. According to the ARG-ST Mapping Lexical Principle for Prepositions in Figure 12.2 the preposition *in* ('in') can take one nominal argument with one unrealized complement. Thus the syntactic and semantic properties of this complement are determined not by the preposition but by the noun. Both the noun and its unre-

alized complement are mapped to the COMPS list of *in* ('in'), and, according to the constraints on the head-complement-structures for prepositions, they are syntactically selected by *in* ('in').

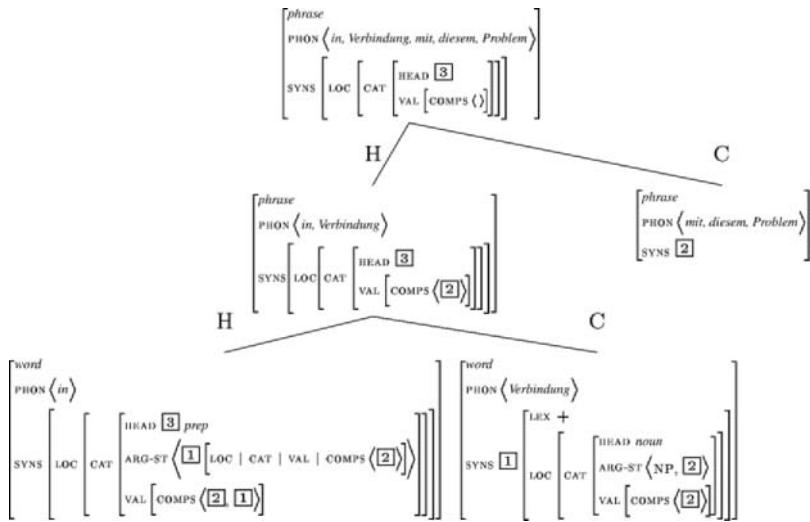


Figure 12.3. The structure of the PP in *Verbindung mit diesem Problem*

The empirical observations of Section 1.1 can be explained by this analysis. The first complement selected by *in* ('in') is the lexical noun. Restrictive adjectives or modifying PPs are both specified as combining only with complement-saturated nouns. Thus, adjunction to complement-unsaturated nouns is blocked. The same restriction holds for determiners and quantifiers in German. These constraints, existing in the grammar independently of the principles of the CPP's syntax, explain the apparent lexical fixedness of the  $P_1N_1$  sequences (cf. 1) and (2) without additional stipulations. The combination *in Verbindung* ('in connection') selects the complement of the noun as its own complement, forming a PP.

Exactly the same lexical entry for *in* ('in') and the same set of principles license PPs headed by non-raising prepositions, such as the PP *in einer engen Verbindung mit den Beratern* ('in a close connection with the advisers').

## 2. Semantic Aspects

In the previous section we have argued that the syntactic structure of CPPs consisting of  $P_1 N_1 P_2 NP$  sequences can be described by use of the raising mechanism which enables prepositions to raise and syntactically realize complements of their arguments. These expressions are thus licensed by virtue of

regular principles of syntax. One may therefore expect that the meaning of these PPs is an instance of regular compositional semantics. We will demonstrate that this is indeed the case, adopting the semantic framework of *Lexicalized Flexible Ty2* (LFTy2; (Sailer, 2003)). In this section we will first present LFTy2 and then show how the meaning of CPPs can be computed on the basis of our syntactic assumptions.

## 2.1 Lexicalized Flexible Ty2

LFTy2 is an adaptation to HPSG of *Flexible Montague Grammar* (Hendriks, 1993). We will take the CONTENT value of a sign to be an expression of a standard semantic representation language, in this case Ty2 (Gallin, 1975). Lexical elements are assigned an expression of Ty2 as their *basic translation*. The CONTENT value of a phrase is the functional application of the CONTENT values of the daughters. In addition, flexible semantic systems provide a number of type shifting operations. These are needed to make the semantic types of sisters compatible with each other, for scope ambiguities and for coordination (see (Hendriks, 1993)). In accordance with (Bouma, 1994) we will apply shifting only to lexical elements. As an illustration, see the PP in (8).

(8) Peter schlief in einem Hotel.  
 Peter slept in a hotel  
 $\exists x[\text{hotel}(x) \wedge \exists e[\text{in}(x, e) \wedge \text{sleep}(e, p)]]$

The semantic derivation of the PP is outlined in Figure 12.4. Every word is assigned a basic translation. The logical form of the NP *einem Hotel* ('a hotel') results from functional application. Since this logical form is of type  $(et)t$  it cannot immediately combine with the basic translation of *in* ('in'),  $z_e R_{e(et)} \cdot \exists e[\text{in}(z, e) \wedge R(u)(e)]$ . LFTy2 offers a shifting operation, called AR (*argument raising*), which raises the type of a semantic argument. Here the first semantic argument of *in* is raised to the type  $(et)t$  in order to be compatible with the NP.<sup>4</sup>

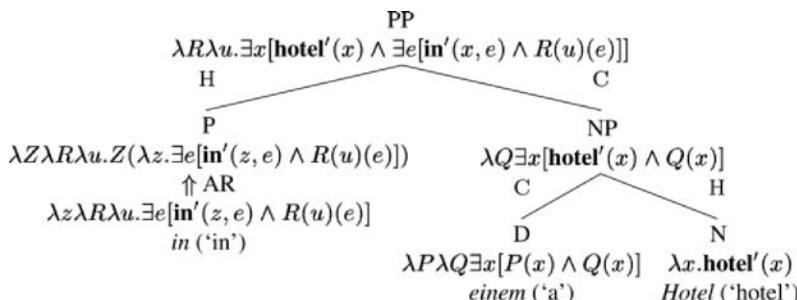


Figure 12.4. The structure of the PP *in einem Hotel* ('in a hotel')

Verbal complexes are the prototypical examples for raising structures, i.e. semantic arguments are not realized as the syntactic complements of the selecting item. Since we plan to analyse CPPs syntactically in analogy with verbal complexes in German, we will first sketch the semantic analysis for verbal complexes. We will then demonstrate that this analysis carries over directly to the PP data. To illustrate this, Figure 12.5 indicates the syntactic structure and the semantic derivation of the VP Fido füttern will (Fido feed want, ‘want to feed Fido’).<sup>5</sup>

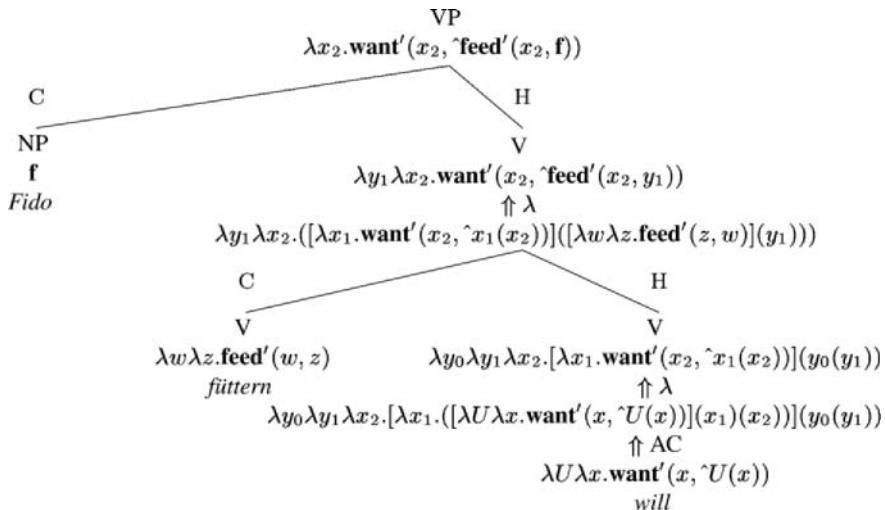


Figure 12.5. The structure of the VP Fido füttern will (Fido feed want, ‘wants to feed Fido’)

The LFTy2 fragment in (Sailer, 2003) does not account for syntactic argument raising. In (??) a shifting operation, AC (*argument composition*), is introduced to achieve the correct identification of syntactic constituents and their semantic roles in raising structures. The definition states that if a functor takes an argument of a certain type  $a_i$ , it can then combine alternatively with a number of other arguments, which also combine to form an expression of type  $a_i$ .

Argument Composition (AC):

AC is a relation between two expressions  $\alpha$  and  $\beta$  such that if  $\alpha$  is of some type  $a_1(\dots(a_i(\dots(a_n b)\dots)\dots)$ , then  $\beta$  is some term  $x_1 \dots x_{(i-1)} y_0 y_1 \dots y_m x_{(i+1)} \dots x_n \cdot [(x_i \cdot \alpha(x_i) \dots (x_n))(y_0(y_1) \dots (y_m))]$  where each  $x_j$  is of type  $a_j$ ,  $y_0$  is of some type  $c_1(\dots(c_m a_i)\dots)$ , and each  $y_k$  is of type  $c_k$ .

In Figure 12.5,  $\alpha$  is the basic translation of will (wants’). For clarity, we have used exactly the same variable names as in the definition of AC. The first

semantic argument of will ('wants') determines  $a_i = et$ .  $y_0$  has the type of the bare infinitival complement *füttern* ('feed'),  $e(et)$ . The direct object of *füttern* ('feed') is syntactically raised and, consequently, its semantic counterpart  $y_1$  appears as an extra argument of type  $e$  in the type-shifted expression. This new expression combines with the basic translation of the verb *füttern* ('feed'). As desired,  $y_0$  combines with  $y_1$  to form an expression of type  $a_i = et$ . For clarity we have indicated the resulting expression before and after  $\lambda$ -conversion ( $\uparrow \lambda$ ).

## 2.2 The Meaning of CPPs

We can now address the interpretation of CPPs. We will show that the syntactic structures assumed for these combinations can be interpreted compositionally. To illustrate this we will examine the PP *in Verbindung mit x* ('in connection with x'). We will argue that the words in this combination occur with a logical form which is also available in other combinations, and that the logical form of the overall PP results from the application of shifting operations and functional application as discussed in the previous subsection.

The preposition *in* ('in') occurs in the PP *in Verbindung mit x* ('in connection with x') with its metaphorical non-spatial meaning, just as in many other combinations (cf. 9). For our purpose, we simply assume the same basic translation of *in* ('in') as in Figure 12.4. The preposition *mit* ('with') is used as a selected preposition. Therefore, it does not contribute an independent meaning and is translated as the identity function ( $\lambda x. x$ ). It occurs with this translation in other combinations as well, such as *mit Fisch handeln* (with fish deal, 'to deal in fish').

(9) in einer Beziehung/ einer Relation/ diesem Zusammenhang  
in a connection/ a relationship/ this context

The noun *Verbindung* ('connection') is a nominalization of the verb *verbinden* ('connect'). The basic translation of the verb is *zyxe.connect(e,x,y,z)*, where  $e$  is a "connecting" eventuality, in which  $x$  connects  $y$  with  $z$ . In an HPSG account of -ung-nominalizations in German, (Reinhard, 2001) proposes that the suffix -ung raises the arguments of the verbal base with which it combines. Which of these arguments can be realized and how they can be realized in syntax depends on the verb class. The example in (10) shows different possibilities of syntactic argument realization.

(10) Eine Verbindung (von bin Laden) mit Hussein wäre absurd.  
a connection of bin Laden with Hussein would be absurd  
'A connection (of bin Laden) with Hussein would be absurd.'

In (10) the underlying subject of *verbinden* ('connect') remains unexpressed. The underlying direct object is also optional. Unrealized arguments are semantically present but unspecified. Thus we assume that they are ex-

istentially bound within the noun's logical form. In (10) this can result in a content value of the form  $ze. \exists x \exists y [connect(e, x, y, z)]$ .

Finally, the PP in (11) has no determiner. The absence of a determiner also has the effect of existential quantification. For further combinatorics, the NP with no determiner must be of type  $(et)t$ . Thus existential quantification over the referential argument leads to a logical form similar to that of a quantified NP, i.e. to the expression  $zP. \exists e [\exists x \exists y [connect(e, x, y, z)] \Lambda P(e)]$  in (11).

(11) Die Raumfähre flog mehrere Tage ohne Verbindung mit der Bodenstation durchs All.

the space shuttle flew several days without connection with the ground station through the space

'For several days the space shuttle flew through space without connection to the ground station.'

This is exactly the logical form we need for the interpretation of the CPP in Verbindung mit x ('in connection with x'). The syntactic structure and the semantic derivation are shown in Figure 12.6. Both are parallel to what is depicted for the VP in Figure 12.5. The basic translation of in ('in') first undergoes AR in order to be of the appropriate type to combine with a quantified NP. Then AC is applied and the resulting expression has two semantic argument ( $y_0$  of type  $e(et)t$  and  $y_1$  of type  $e$ ) instead of the single semantic argument  $V$  of type  $(et)t$  in the input to AC. This demonstrates that the meaning of the PP can be computed on the basis of independently motivated meaning assignments and shifting rules.

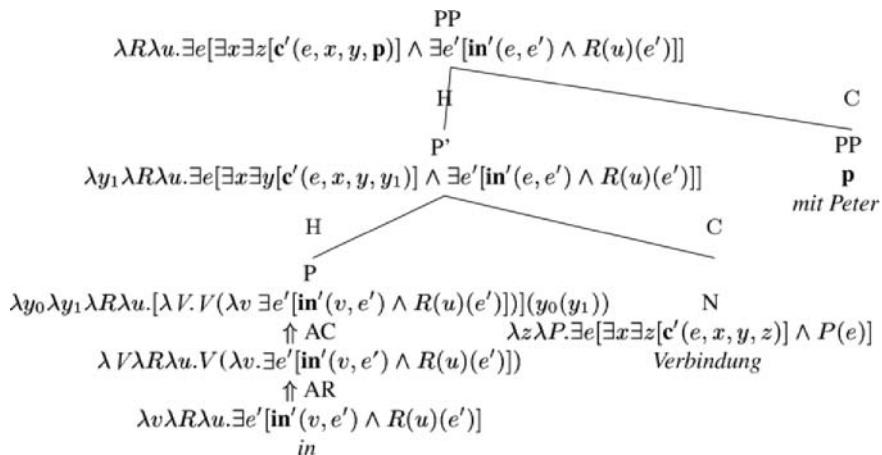


Figure 12.6. The structure of the PP in Verbindung mit Peter ('in connection with Peter')

One can treat most  $P_1 N_1 P_2 NPs$  in which  $N$ s are deverbal event nominalizations as being semantically regular, i.e. licensed by regular translations and

regular derivational operations. This semantic and syntactic regularity explains their high productivity in contemporary German. In the next section we will discuss two types of irregular  $P_1N_1P_2NP$  combinations.

### 3. Irregular Combinations

In this section we will discuss subtypes of CPPs which behave differently with respect to the two *regularity properties* in (??) which are adopted from (Sailer, 2003).<sup>6</sup> If a given  $P_1N_1P_2NP$  sequence lacks at least one of these properties, we will consider it irregular, i.e. of idiomatic character.

**RP1:** Every element of the PP can be attributed a meaning with which it also occurs independent of the combination under consideration.

**RP2:** The meaning of the entire PP is arrived at by combining the meanings of its parts in a regular way.

If we reconsider the analysis of *in Verbindung mit x* ('in connection with *x*') in Figure 12.6, we see that this CPP shows both regularity properties. Firstly we argued that all the lexical elements in the combination appear with the same meaning assignment in other structures (RP1). Secondly we applied only the rules of syntactic and semantic combination which are independently required in the language (RP2).

Whereas *in Verbindung mit x* ('in connection with *x*') can be described as a fully regular combination, the following two subsections will be devoted to  $P_1N_1P_2NP$  combinations which show irregularities with respect to RP1 or RP2. Nevertheless, there are approaches which provide the prerequisites to account for these combinations: external selection (Soehn, 2003) and phrasal lexical entries (Sailer, 2003). We will outline both approaches and show how to apply them to account for the more idiosyncratic CPPs.

#### 3.1 Bound Words

In some irregular  $P_1N_1P_2NP$  sequences the  $N_1$  is a so-called *bound word*, e.g. *in Anbetracht von x* ('in consideration of *x*'). The entire PP is semantically decomposable, and thus satisfies the condition of semantic regularity in RP2. However, RP1 has not been satisfied, since not all components of that PP may occur with the same meaning in other contexts: the noun *Anbetracht* ('consideration') can only occur in combination with the preposition *in* ('in').

To account for bound words within PPs in general, (Soehn, 2003) generalizes the external selection mechanisms of HPSG (cf. the MOD and SPEC features). (Soehn, 2003) assumes that in every type of phrase the non-head daughter can determine syntactic and semantic properties of the head daughter. This idea is realized by conflating the attributes MOD and SPEC into one attribute XSEL (*external selection*), which is appropriate for the sort *head* and takes a *synsem* object as its value. In addition the so-called PRINCIPLE OF

EXTERNAL SELECTION(PXS) ensures the identity of the XSEL value of the non-head and the SYNSEM value of the head, similar to the SPEC-PRINCIPLE, which has become obsolete.

In the lexical entry of *Anbetracht* ('consideration') in Figure 12.7, the XSEL value is specified as a synsem object with [pform in]. This specification and the PXS will ensure the occurrence of *Anbetracht* ('consideration') exclusively within a PP headed by the preposition *in* ('in'). For freely occurring words, the XSEL value is underspecified.<sup>7</sup>

The PP *in Anbetracht von x* ('in consideration of *x*') is a  $P_1N_1P_2NP$  expression. Therefore, the xsel value of *Anbetracht* ('consideration') explicitly requires the preposition *in* ('in') to raise the argument of the bound word, i.e. the PP *von x* ('of *x*') (2 in the figure). The lexical entry for the noun *Anbetracht* in Figure 12.7 shows that we can smoothly merge the external selection approach of (Soehn, 2003) with the complement raising approach.

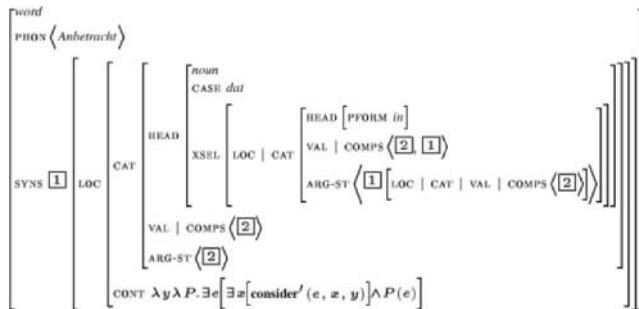


Figure 12.7. The relevant part of the lexical entry of the noun *Anbetracht* ('consideration')

Assuming the usual non-spatial meaning for *in* ('in') selecting *Anbetracht* ('consideration'), we can derive the meaning of the entire PP parallel to the derivation in Figure 12.6. This shows that we can smoothly merge the external selection approach of (Soehn, 2003) with the complement raising approach.

### 3.2 Phrasal Lexical Items

There are also  $P_1N_1P_2NP$  expressions which escape a compositional treatment, such as *an Hand von x* (*at hand of x*, 'by means of *x*'), *an Stelle von x* (*at place of x*, 'in lieu of *x*') or *auf Grund von x* (*on base of x*, 'by virtue of *x*'). This type is significantly less frequent in German than the fully regular combinations. These expressions consist of lexical entities of which each one also appears outside the particular PP. When considering the meaning of any of these PPs it is highly problematic to assign a combination-specific meaning to its particular elements such that the meaning of the entire PP could be de-

rived compositionally. Therefore these combinations do not exhibit RP2. This irregular behavior makes the assumption plausible that these expressions are licensed directly by the lexicon. In this subsection we will provide an analysis of this type of CPPs using the expression *an Hand von* (*at hand of*, 'by means of') as a prototypical example.

In the architecture of (Pollard and Sag, 1994) all syntactically complex signs, i.e. all phrases, are subject to the regular principles of syntactic and semantic combination. However, idiomatic expressions of the type kick the bucket ('die') cannot be handled with this kind of approach. To overcome this empirical deficiency, (Sailer, 2003) uses *Phrasal Lexical Entries* (PLEs).<sup>8</sup>

(Sailer, 2003) introduces an attribute *coll* (*context of lexical licensing*) on the sort sign. Signs which are directly licensed by the lexicon have the specification *[coll +]*, whether they are words or phrases. On the other hand, signs which are licensed by ID schemata or lexical rules have the specification *[coll -]*. Consequently there is a **LEXICON PRINCIPLE** which lists the lexical entries for all signs with a positive *coll* value. This principle contains the usual lexical entries for words (*LE<sub>i</sub>*) as well as phrasal lexical entries for idiosyncratic phrases (*PLE<sub>i</sub>*). Additionally the antecedents of principles of regular combination, such as the **ID PRINCIPLE** and the **SEMANTICS PRINCIPLE**, are restricted to phrases with a *[coll -]* specification.<sup>9</sup>

The **LEXICON PRINCIPLE**:

$$\left[ \begin{matrix} \text{sign} \\ \text{COLL +} \end{matrix} \right] \longrightarrow LE_1 \vee \dots \vee LE_n \vee PLE_1 \vee \dots \vee PLE_m$$

We can apply this approach to PPs such as *an Hand von x* (*at hand of x*, 'by means of *x*'). We assume a PLE for the combination *an Hand* (*at hand*, 'by means') which requires a genitive NP or a von-PP as its complement. This PLE is outlined in Figure 12.8. It is important to note that even though the phrase *an Hand* (*at hand*, 'by means') is irregular, its daughters *an* ('at') and *Hand* ('hand') occur as exactly the same words in other contexts. However, the semantic contributions of the words are not combined to form the content of the phrase. Instead, the phrase as a whole receives an idiosyncratic meaning.

The use of the phrase *an Hand* (*at hand*, 'by means') in larger structures is illustrated in Figure 12.9. Note that the *COLL* values of the phrases *von Prolog* ('of Prolog') and *an Hand von Prolog* (*at hand of Prolog*, 'by means of Prolog') are specified as *-*, since these phrases are licensed by the regular constraints of grammar. In contrast, the *COLL* value of the phrase *an Hand* (*at hand*, 'by means') is specified as *+*. As an internally irregular expression, the phrase *an Hand* (*at hand*, 'by means') is licensed immediately by the lexicon.

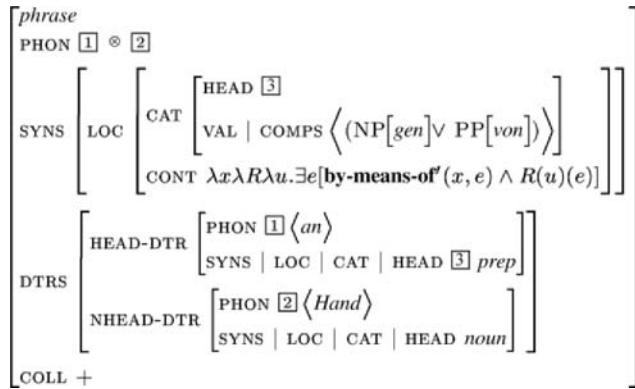


Figure 12.8. Outline of the phrasal lexical entry of an Hand (*at hand*, 'by means')

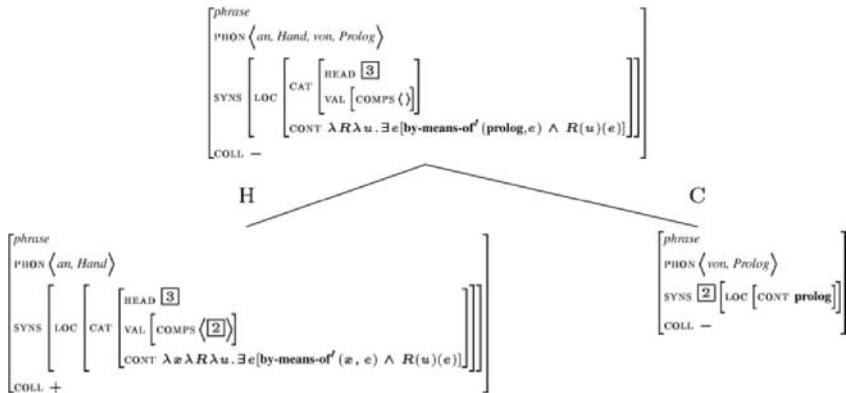


Figure 12.9. The structure of the PP *an Hand von Prolog* (*at hand of Prolog*, 'by means of Prolog')

In this section we have demonstrated that our account of CPPs interacts in an empirically adequate way with HPSG approaches to irregularity phenomena such as the xsel approach to distributional idiosyncrasies and the phrasal lexical entry approach to combinatorial irregularities.

## 4. Summary

(Trawiński, 2003) discusses syntactic properties of  $P_1 N_1 P_2 N_2$  sequences which are the basis for complement raising analysis. Based on this analysis, we have investigated further properties of these CPPs focusing on semantic aspects. The objective of our investigations was to examine whether the semantic representation of these expressions can be derived compositionally. We

have thereby seen that the expressions discussed are not uniform with regard to their semantic behavior, forming three classes: CPPs which can be analyzed compositionally (*in Verbindung mit x* ('in connection with *x*')), CPPs involving bound words which can also be treated within the combinatorial semantics but which require some mechanism to describe distributional properties of the particular bound words (*in Anbetracht von x* ('in consideration of *x*')), and CPPs which cannot be handled by virtue of common derivational methods (*an Hand von x* (*at hand of x*, 'by means of *x*')). However, we have shown that the available HPSG approaches, i.e. the external selection approach of (Soehn, 2003) and the phrasal lexical entry approach of (Sailer, 2003), provide the necessary means to account for all of these data.

## Acknowledgments

We would like to thank Frank Richter and the anonymous reviewers for their helpful comments. We are also grateful to Guthrun Love for her help with the proofreading in English.

## Notes

1. Cf. (Lindqvist, 1994), (Quirk and Mulholland, 1964), (Beneš, 1974), etc.
2. It is unclear how many  $P_1 N_1 P_2$  expressions there are in German. (Schröder, 1986) identifies more than 90. (Beneš, 1974) itemizes 160 examples, thereby emphasizing the incompleteness of his list. In any case, these word combinations do not form a marginal class of expressions in contemporary German. For discussion on CPPs in German see also (Meibauer, 1995).
3. For further applications of the raising mechanism within the HPSG grammar framework see e.g. (Hinrichs and Nakazawa, 1989), (Hinrichs and Nakazawa, 1994), (Meurers, 2000) or (De Kuthy, 2000).
4. We deliberately simplify the treatment of the eventuality variable *e* when we assume that the quantifier which binds *e* is introduced by the preposition. This simplification has no bearing on the main argumentation in this paper.
5. We have left out the eventuality variables in this example for simplification.
6. (Sailer, 2003) applies analogous criteria to VPs. There, expressions with *bound words* such as make headway, violate the first regularity property. Non-decomposable idiomatic expressions such as kick the bucket show the corresponding violation of the second property.
7. Nouns often show idiosyncratic preferences for a particular preposition, such as *in/\* bei Verbindung mit* ('in / \*at connection with'). The xsel approach can capture this. For the lexical entry of Verbindung ('connection') we only have to add a constraint stating that if the xsel value of the noun is a raising preposition, then this preposition has the pform value in ('in'). The same solution can be applied to the noun-specific choice of support verbs.
8. For an alternative constructional approach to idioms see (Riehemann, 2001) or (Sag et al., 2002).
9. For phenomena discussed in this paper it is sufficient to adopt the simplified usage of the *COLL* attribute as presented in Section 8.1 of (Sailer, 2003). In his Section 8.3 (Sailer, 2003) assumes that the *COLL* attribute takes a list of signs as its value, such that for every non-lexical sign, the *COLL* value is an empty list. The *coll* value of a lexical sign is a singleton list containing the root sign of the utterance in which this lexical sign occurs. With this more complex mechanism, CPPs with bound words can also be described. However, as elaborated in (Soehn and Sailer, 2003), this more general use of *coll* might be too powerful.

## References

Beneš, E. (1974). Präpositionswertige Präpositionalfügungen. In Engel, U. and Grebe, P., editors, *Sprachsystem und Sprachgebrauch. Festschrift für Hugo Moser zum 65. Geburtstag. Teil I.*, number 33 in Sprache der Gegenwart, pages 33–52. Schwann, Düsseldorf.

Bouma, G. (1994). Calculated Flexibility. In Bunt, H., Muskens, R., and Ren-tier, G., editors, *Proceedings of the International Workshop on Computational Semantics*, pages 32–40. Katholieke Universiteit Brabant.

De Kuthy, K. (2002). *Discontinuous NPs in German — A Case Study of the Interaction of Syntax, Semantics and Pragmatics*. CSLI Publications, Stanford.

Fries, N. (1988). *Präpositionen und Präpositionalphrasen im Deutschen und im Neugriechischen*. Number 208 in Linguistische Arbeiten. Max Niemeyer Verlag, Tübingen.

Gallin, D. (1975). *Intensional and Higher-Order Modal Logic*. North-Holland, Amsterdam.

Hendriks, H. (1993). *Studied Flexibility*. ILLC Dissertation Series 1995-5. Institute for Logic, Language and Computation, Amsterdam.

Hinrichs, E. and Nakazawa, T. (1989). Flipped Out: Aux in German. In *Papers from the 25th Regional Meeting of the Chicago Linguistic Society*, pages 193–202, Chicago, Illinois.

Hinrichs, E. and Nakazawa, T. (1994). Linearizing AUXs in German Verbal Complexes. In Nerbonne, J., Netter, K., and Pollard, C., editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in CSLI Lecture Notes, pages 11–37. CSLI Publications, Stanford.

Lindqvist, Ch. (1994). *Zur Entstehung der Präpositionen im Deutschen und Schwedischen*. Max Niemeyer Verlag, Tübingen.

Meibauer, J. (1995). Komplexe Präpositionen — Grammatikalisierung, Metapher, Implikatur und *Division of Pragmatic Labour*. In Liedtke, F., editor, *Implikaturen. Grammatische und pragmatische Analysen*, number 343 in Linguistische Arbeiten, pages 67–74. Max Niemeyer Verlag, Tübingen.

Meurers, W. D. (2000). *Lexical Generalizations in the Syntax of German Non-Finite Constructions*. PhD thesis, Universität Tübingen. Published as: Arbeitspapiere des SFB 340, Nr. 145.

Pollard, C. J. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Quirk, R. and Mulholland, J. (1964). Complex Prepositions and Related Sequences. In *English studies presented to R. W. Zandvoord on the occasion of the 70th birthday, Supplement to Vol. 45*, pages 64–73, Amsterdam.

Reinhard, S. (2001). *Deverbale Komposita an der Morphologie-Syntax-Semantik-Schnittstelle: ein HPSG-Ansatz*. PhD thesis, Universität Tübingen.

Riehemann, S. Z. (2001). *A Constructional Approach to Idioms and Word Formation*. PhD thesis, Stanford University.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*. Springer-Verlag, Heidelberg/Berlin.

Sailer, M. (2003). *Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar*. PhD thesis (2000), Universität Tübingen. Published as: *Arbeitspapiere des SFB 340*, Nr. 161.

Schröder, J. (1986). *Lexikon deutscher Präpositionen*. Verlag Enzyklopädie, Leipzig.

Soehn, J.-P. (2003). *Von Geisterhand zu Potte gekommen. Eine HPSG-Analyse von PPs mit unikaler Komponente*. Master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft.

Soehn, J.-P. and Sailer, M. (2003). At First Blush on Tenterhooks. About Selectional Restrictions Imposed by Nonheads. In Jäger, G., Monachesi, P., Penn, G., and Wintner, S., editors, *Proceedings of FGVienna: The 8th Conference on Formal Grammar*, pages 149–161. To appear also as CSLI Publications Online Proceedings.

Trawiński, B. (2003). Licensing Complex Prepositions via Lexical Constraints. In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 97–104, Sapporo, Japan.

Trawiński, B. (2003). A New Application for Raising in HPSG: Complex Prepositions. In Jäger, G., Monachesi, P., Penn, G., and Wintner, S., editors, *Proceedings of FGVienna: The 8th Conference on Formal Grammar*, pages 163–175. To appear also as CSLI Publications Online Proceedings.

## Chapter 13

# DISTRIBUTIONAL SIMILARITY AND PREPOSITION SEMANTICS

Timothy Baldwin

*CSLI, Stanford University*

[tbaldwin@csli.stanford.edu](mailto:tbaldwin@csli.stanford.edu)

**Abstract** Prepositions are often considered to have too little semantic content or be too polysemous to warrant a proper semantic description. We illustrate the suitability of distributional similarity methods for analysing preposition semantics by way of an inter-preposition similarity task, and make the claim that any semantic account of preposition semantics must be partially conditioned on valence.

**Keywords:** distributional similarity, latent semantic analysis, preposition semantics, Roget's thesaurus, preposition valence

### 1. Introduction

While nouns, verbs and adjectives have received considerable attention in terms of both lexical semantic language resource development Ike-  
hara:1991,Mahesh:1996,Fellbaum:1998 and automatic ontology construction (Grefenstette, 1994), (Lin, 1998a), (Widdows et al., 2002), relatively little work has been done on creating resources for prepositions. Perhaps a large part of the reason for this is that the semantics of a transitive preposition can be bleached and determined largely by the semantics of the head noun it governs (e.g. at last, on Wednesday, in question: (Pustejovsky, 1995)) or its governing verb (e.g. refer to, talk about). However, many prepositions also have predicative usages (e.g. time is up, the cheese is off, flairs are in), and the semantics of peripheral PPs is determined largely by the preposition (e.g. from March, in Toulouse, by the gate, at/in Stanford). Accordingly, some account of preposition semantics seems unavoidable.

There is a relative sparsity of computational research on preposition semantics, which can perhaps be explained by the perception that prepositions are

both semantically vacuous and distributionally highly promiscuous, and consequently have a very low information content. This is most pronounced in bag-of-words tasks such as information retrieval where prepositions are generally listed in “stop word” lists for exclusion as index terms.

Our interest is in testing the viability of distributional methods in the derivation of a model of preposition semantics, working under the hypothesis that preposition semantics are stable enough that they can be classified accurately by distributional similarity techniques. Our approach here is based on the distributional hypothesis of (Harris, 1968) that similar words tend to occur in similar linguistic contexts. This observation has been used to explain various aspects of human language processing, from lexical priming (Lund, 1995) to retrieval in analogical reasoning (Ramscar, 2003). It has also been employed in a range of natural language processing tasks, including word sense disambiguation (Schütze, 1998) and automatic thesaurus construction (Lin, 1998b). To our knowledge it has not previously been used to analyse the meaning of closed-class words.

As well as demonstrating the ability of similarity methods to capture intuitive correlations in the semantics of prepositions, we are interested in unearthing semantic anomalies between particles and transitive prepositions and motivating a valence-conditioned classification of English prepositions. Intransitive prepositions (Huddleston et al., 2002) (which we will interchangeably refer to as particles) are valence-saturated and occur most commonly as: (a) components of larger multiword expressions (notably verb particle constructions, or VPCs, such as *pick up*, *call in* and *chicken out*), (b) predicates (e.g. *time is up*, *flairs are in*) or (c) prenominal modifiers (e.g. *the up escalator*, *off milk*). Transitive prepositions, on the other hand, select for NP complements to form prepositional phrases (PPs, e.g. *at home*, *in the end*). The bare term preposition is valence-underspecified. Hereafter, we will index intransitive prepositions with the suffix “” (e.g. *up*) and transitive prepositions with the suffix “” (e.g. *up*) in cases where we wish to refer to a particular valence.

It is relatively easy to find senses which are attested for only intransitive prepositions (e.g. the *hip/in* fashion sense of *in* above) and also uniquely transitive prepositions (e.g. *from*) which by definition do not have intransitive semantics. Of greater interest is the degree of correlation between intransitive and transitive preposition sense according to automatically-derived semantic classifications. That is, we seek to quantify the degree of semantic divergence between intransitive and transitive usages of different prepositions.

One piece of preliminary evidence which underlines the potential applicability of the distributional hypothesis to prepositions comes from the field of English part-of-speech (POS) tagging. All major POS tagsets<sup>1</sup> prefer to underspecify valence (e.g. there is no tag distinction between intransitive and transitive verbs), with the glaring exception of prepositions which are in all cases

partitioned into intransitive and transitive instances. If there were a sharp demarcation in wordform between intransitive and transitive prepositions in English, this finding would perhaps not be surprising. However, a summary analysis of the written component of the British National Corpus (BNC, (Burnard, 2000)) reveals that while the type overlap between the two classes is only around 8%, the token overlap is roughly 70%. That is, roughly 70% of preposition token instances are potentially ambiguous between an intransitive and transitive usage. Given that taggers are able to deal with this ambiguity, generally using the immediate lexical context of a given preposition token, it would appear that intransitive and transitive usages of a given preposition are to some degree distributionally dissimilar. In this paper, we seek to confirm that this distributional dissimilarity correlates with semantic disparity, and at the same time determine whether semantically-related prepositions are distributionally similar.

The remainder of this paper is structured as follows. sec:preposition-similarity outlines our implementation of distributional similarity as a means of modelling simplex preposition semantics. sec:gold-standard describes the two gold standard sources of English preposition similarity used in this research. sec:evaluation presents quantitative and qualitative evaluation of our method. We outline related research in sec:related-research and conclude the paper in sec:conclusion.

## 2. Calculating inter-preposition similarity

In this paper, we consider the task of inter-preposition similarity, that is determination of the relative similarity of different preposition pairs. The procedure used to calculate preposition similarity is knowledge-free and based on Latent Semantic Analysis (LSA, (Deerwester et al., 1990)). Our technique is very similar to the approach taken to building a “context space” by (Schütze, 1998). We measured the frequency of co-occurrence of our target words (the 20,000 most frequent words), with a set of 1000 “content-bearing” words (we used the 51st to the 1050th most frequent words, the 50 most frequent being taken to have extremely low infomation content). A target word was said to co-occur with a content word if that content word occurred within a window of 5 words to either side of it. In order to overcome data sparseness, we used Singular Value Decomposition (SVD) to reduce the dimensionality of the feature space from 1000 to 100. This limits each target word vector to 100 factors which reflect the patterns of association in the matrix, allowing relations to be discovered between target words even if there is not direct match between their context words. We used the various tools in the GTP software package, created at the University of Tennessee<sup>2</sup>, to build these matrices from the co-occurrence data and to perform SVD analysis.

The resulting representation is a 100-feature vector for each target word. Using this we can calculate the similarity between two terms by finding the cosine of the angle between their vectors.

As mentioned above, we distinguish prepositions according to valence, and seek to provide evidence for divergences in transitive and intransitive preposition semantics. This is achieved according to Methods prep1 and prep2, as detailed below. We evaluate the methods over the written component of the BNC (90m words).

**Method PREP1** First, we ran the above method over wordforms. With this method, we are thus unable to differentiate intransitive and transitive usages of a given preposition.

**Method PREP2** Second, we ran our method including POS tags from the output of the RASP system Briscoe:Carroll:2002, i.e. treating each wordform-POS tag pair as a single token. The RASP tagger is based on the CLAWS-4 tagset, and thus offers a fine-grained distinction between different kinds of prepositions and particles. In extracting our context space we collapsed the different varieties of prepositions to give us one category for transitive prepositions and one for intransitive prepositions.

While LSA is generally applied simply to wordforms, we are certainly not the first to integrate POS tags with the wordforms to generate POS-sensitive semantic models. e.g. (Widdows, 2003) demonstrated the superiority of POS-conditioned semantic models on a taxonomy induction task.

### 3. Gold standard sources of inter-preposition similarity

In order to evaluate the quality of the preposition similarities derived via LSA, we turn to the only two large-scale public-domain resources we are aware of that provide a unified, systematic account of preposition semantics: the LCS-based preposition lexicon of (Dorr, 1997)<sup>3</sup> and the 1911 edition of Roget's thesaurus.<sup>4</sup>

#### 3.1 LCS-based preposition lexicon

The preposition lexicon of (Dorr, 1997) is couched in lexical conceptual semantics Jackendoff85, and is made up of 165 English prepositions classified into 122 intransitive and 375 transitive senses. Each preposition sense is described in the form of an LCS-based representation such as (toward Loc (nil 2) (UP Loc (nil 2) (\* Thing 6))), corresponding to the up the stairs sense of up. (Resnik and Diab, 2000) propose a method for deriving similarities from LCS representations by: (1) decomposing them into feature sets, (2) calculating the information content  $I(f)$  of each unit feature  $f$  based on the

overall feature distribution, and (3) measuring the similarity between two LCS representations according to:

$$(13.1) \quad sim_{LCS}(e_1, e_2) = \frac{2 \times I(F(e_1)) \cap I(F(e_2))}{I(F(e_1)) + I(F(e_2))}$$

where  $e_1$  and  $e_2$  are lexicon entries,  $F(e_i)$  is the decomposed feature set associated with  $e_i$ , and  $I(F(e_i))$  is the information content of that feature set. (Resnik and Diab, 2000) define the similarity between two words to be the maximum value of  $sim_{LCS}(e_1, e_2)$  over the cross product of all lexical entries for the words.

One key feature of this lexicon is that it captures the transitive and intransitive preposition senses separately, but within a common representation. As a result, we are able to derive similarities (a) at the wordform level, comparing all senses of a given preposition pair irrespective of valence, and (b) in a valence-sensitive fashion, calculating  $sim_{LCS}$  only for lexicon entries of equivalent valence. This facilitates independent analysis of the correlation of prep1 (wordform-based) and prep2 (POS-conditioned) with the preposition lexicon-derived similarities.

It is worth pointing out that, in the context of an experiment testing correlation with human judgements on verb similarity, (Resnik and Diab, 2000) found  $sim_{LCS}$  to be inferior to a number of taxonomic similarity measures and a distributional similarity measure. It is thus with a certain degree of reservation that we reimplement their method, noting however that the taxonomic similarity avenue is not open to us due to the absence of a taxonomy.

### 3.2 Roget's thesaurus

The 1911 edition of Roget's thesaurus incorporates around 100K lexical entries in a total of 1000 semantic classes. In the original 1911 configuration the classes have no explicit relational structure, although subsequent work has been done to add hierarchical structure to the thesaurus (e.g. (Kirkpatrick, 1988)). We justify our use of the 1911 edition of Roget's thesaurus on the grounds that (a) there are no restrictions on the use of this version of the thesaurus, and (b) the classification of prepositions is largely unchanged in more recent editions of the thesaurus.

One attraction of Roget's thesaurus is that, within each semantic class, it lists words according to the four basic word classes of noun, verb, adjective and adverb.<sup>5</sup> Because of this cross-listing, we can preserve the experimental setup described above for LSA, calculating inter-preposition similarity either according to wordform or conditioned on POS.<sup>6</sup>

In Roget's thesaurus, prepositions are listed as either adjectives or adverbs, which would superficially appear to correspond to transitive and intransitive prepositions, respectively. In practice, adjectival entries are restricted to pred-

icative and attributive particles such as the *in* crowd and hence limited in number, whereas adverbial entries represent a mix of intransitive and transitive usages. Consider the preposition *up*, for example, which is listed twice as an adjective (Bubble, as in frothy, and Excitation, as in stung to the quick<sup>7</sup>) and twice as an adverb (Height, as in aloft, and Verticality, as in on end). Here, it is not clear whether the adverbial entries are intended to be intransitive, transitive or both. In some cases, the valence is self-evident as the preposition in question is either uniquely transitive (e.g. *from*, listed under Motive) or uniquely intransitive (e.g. *aback*, listed under Rear). Alternatively, a sense may be particular to a given valence. However, more often than not, we have no reliable way of determining the intended valence of each preposition entry. Thus, we are able to make use of the adjectival entries in modelling particle sense, but have no immediate means of capturing strictly transitive preposition sense. Having said this, for the purposes of evaluation, we consider adjectival preposition entries to be particles and adverbial preposition entries to be transitive prepositions.

Given the lack of hierarchical structure in the 1911 edition of Roget's thesaurus, our options for deriving class-to-class and word-to-word similarities are restricted. The simplest means of deriving class-to-class similarities is to calculate the relative lexical overlap; word-to-word similarities can equivalently be obtained by calculating the degree of overlap in class membership of each word. Unsurprisingly, this naive methodology suffers from acute data sparseness, culminating in the vast majority of class or word pairings being assigned a similarity of 0. In order to overcome this shortcoming, we notice that it is possible to describe a word pairing by way of a bipartite graph with the classes each word occurs in as the opposing vertices. We can then represent class similarities as edges in the graph, and calculate word-to-word similarity according to the maximal bipartite matching (i.e. set of edges such that every vertex is joined to some other vertex) with the highest mean edge score. We initialise each class similarity  $sim_C(i,j)$  to 1 iff  $i = j$  and 0 otherwise, such that in the initial configuration, the bipartite graph method is equivalent to the naive class overlap method. We can now iterate between calculating word-to-word and class-to-class similarities—using a bipartite graph with *words* as vertices and *word similarities* as edges in the class-to-class case—and feed the results of the word-to-word similarity recalculations into class-to-class similarity recalculations, and vice versa.

The net effect of this iterative process is to monotonically propagate the effects of class and word overlap, such that both class and word similarities progressively converge to 1. Our driving motivation in this is essentially to “smooth” similarities and eliminate instances of similarity 0. The stopping condition on the method, therefore, is the condition of there being no class similarity  $sim_C(i,j)$  or word similarity  $sim_W(i,j)$  with value 0. In our experiments, this was generally found to occur on the third iteration.

As we are interested only in preposition similarity (and due to limitations on computational resources), we calculate class-to-class similarities only over those classes which contain one of 54 commonly-occurring prepositions, a total of 78 classes; in calculating word-to-word similarity for non-prepositions contained in the 78 classes, we focus on class membership only over the preposition-containing classes.

In addition to evaluating word-to-word preposition similarity according to simplex preposition entries, we test the use of VPCs as a proxy to situated particle semantics. The method here is identical to that for simplex words, except that we additionally look for occurrences of VPCs as contained in a list of VPC types extracted out of the BNC Baldwin:2002c, and record each such occurrence as an instance of the particle contained therein. That is, we do not distinguish between simplex occurrences of the preposition and occurrences within VPCs. This is not intended to be a general claim about semantic head-*edness* or the relative semantic contribution of the particle in VPCs. Rather we are testing the hypothesis that particles with similar semantics will occur with the same classes of verbs.

Due to the inherent complexity of the similarity calculation, we restrict the number of VPCs by counting the VPCs contained in each class not containing a simplex preposition, and including only those VPCs found in the 200 most heavily VPC-populated classes.

## 4. Evaluation

In this section, we evaluate the LSA-based similarities relative to similarities derived from the LCS lexicon and also Roget's thesaurus.

We measure the correlation between the distributional similarities and both LCS- and thesaurus-derived similarities according to Pearson's  $r$ , as applied to the attested pairings of the nine prepositions about, down, in, off, on, out, over, through and up. We determine the correlation for three distinct datasets: (A) preposition similarity according to prep1 (with underspecification of valence); (B) particle similarity according to prep2; and (C) transitive preposition similarity according to prep2. In the case of (A), therefore, we calculate the distributional similarity of prepositions in the absence of POS information, and likewise do not distinguish between intransitive and transitive prepositions in the LCS lexicon. For (B) and (C), on the other hand, we consider only prepositions of fixed transitivity in both the BNC data and LCS lexicon.

### 4.1 Correlation with LCS-based similarities

The mean  $r$  values relative to the LCS-based similarities are given in tab:lcs-sim for datasets A, B and C. While the values are relatively modest, they provide weak evidence for the ability of LSA to capture preposition semantics.

prep1 (A) all	prep2	
	(B) intransitive	(C) transitive
0.304	0.365	0.386

Table 13.1. Correlation ( $r$ ) between the LCS-based and LSA similarities

	Roget's similarity				
	−valence	−valence <sub>VPC</sub>	+valence	+valence <sub>VPC</sub>	vpc
prep1 (A)	0.004	0.080	—	—	0.097
prep2 (B)	−0.183	0.881	−0.235	0.863	0.805
prep2 (C)	−0.173	−0.287	−0.258	−0.205	−0.197

Table 13.2. Correlation ( $r$ ) between the Roget's thesaurus-based and LSA similarities

Perhaps more importantly, the correlations for the intransitive and transitive preposition similarity tasks ((B) and (C), respectively) are higher than that for the valence-underspecified preposition similarity task (A), at a level of statistical significance (based on the two-tailed  $t$ -test,  $p < .05$ ). This suggests that our model of preposition semantics is more stable when valence is specified, providing tentative support for the claim that preposition semantics are to some degree conditioned on valence.

Recall that we had reservations about the quality of similarities produced with this method, based on the findings of (Resnik and Diab, 2000) over a small-scale verb similarity task. Having said this, the fact that both valence-specified models of distributional similarity were found to correlate more highly than the valence-underspecified model would appear to be significant.

## 4.2 Correlation with Roget's-based similarities

We turn next to Roget's thesaurus and calculate the correlation with similarities derived: (a) independently of valence information for simplex preposition entries (conflating adjectival and adverbial preposition entries: −valence), optionally incorporating semantic classes for VPCs (−valence<sub>VPC</sub>); (b) conditioned on valence information (+ valence), once again optionally incorporating semantics classes for VPCs (+ valence<sub>VPC</sub>); or (c) based only on the VPC entries, without the simplex preposition entries (vpc). The results are presented in tab:roget-sim. Note that we compare prep1 against only the valence-underspecified Roget's similarities as there is no obvious way of combining similarities across the two transitivities for a given preposition. Note also that for the valence-specified models, we always compare like with like, e.g. prep2

(B) is only compared against particle similarities in the correlation analysis with  $+\text{valence}$  and  $+\text{valence}_{\text{VPC}}$ .

We see some interesting results. First, the correlation for prep1 is almost 0 in all cases. That is, in the absence of valence information, the relative similarity values for the different prepositions are nearly randomly distributed. Next, the transitive preposition similarities (the row of prep2 (C)) are negatively correlated in all cases, but relatively low. Thorough error analysis is required to determine the case of this negative correlation, but it is worth noting the differential between the  $r$  values for prep2 (C) and prep1 down each column, indicating that the LSA similarities for valence-underspecified prepositions vary significantly over those for transitive prepositions. Recall that, for the purposes of this evaluation, we are treating adverbial preposition entries in Roget's as transitive prepositions, and it is the similarities for these that we are comparing prep2 (C) against. Given our observations above about the mixed nature of adverbial prepositions, it is perhaps not surprising that no real correlation was found. Finally, prep2 (B) produces remarkably similar results to prep2 (C) for the models which do not make use of the VPC data, but when we add in the VPC classes, we find the correlation to be surprisingly high. The combination of VPC data and valence-underspecified preposition entries returns the highest  $r$  value at 0.881. This compares very favourably with the  $r = 0.901$  and  $r = 0.793$  figures cited as inter-annotator correlation for noun and verb similarity tasks (Resnik, 1995), (Resnik, 2000). Indeed, it provides strong evidence that, at least when viewed in the context of VPC occurrence, particle semantics are well-defined and can be captured effectively by distributional similarity methods.

### 4.3 Discussion of the results

What the above experiments show is, first, that LSA can be applied successfully to the task of inter-preposition similarity modelling. This in itself is a surprising finding, given that the standard practice in established domains for LSA such as information retrieval (IR) is to ignore all prepositions and other stop words. This result is particularly striking as it was validated over heterogeneous sets of similarities, derived from formal semantic representations in the first instance and word clusters in the second.

In our second experiment based on Roget's thesaurus, we found that complementing the simplex inventory of preposition sense led to a huge increase in correlation with the LSA similarities. One could possibly argue that this finding is a by-product of the fact that we are deriving our similarities from Roget's in a similar fashion to LSA, in that we are making use of a context window in calculating the similarities. However, when we consider what role the VPCs are playing in the similarity calculation, it quickly becomes evident that this is

not the case. At no point do we compare which verbs different particles co-occur with. Instead, we take note of which semantic classes those VPCs occur in, and base our similarity calculation on class overlap as per usual. That is, for two particles to be similar, they must combine with verbs (and not necessarily the same verbs) to generate VPCs of the same semantic types. As an illustration of this, consider the particle pairing back and down. In the +valence model,  $sim_W(back_0, down_0) = 0$  as down is not listed as an adjective in Roget's. In +valence<sub>VPC</sub>, on the other hand,  $sim_W(back_0, down_0) = 0.68$  as VPCs such as fall back/back down and keep back/tie down occur in the same semantic classes.

In the first experiment based on the LCS lexicon, we were able to demonstrate modest gains in correlation by conditioning similarity on valence for both transitive and intransitive prepositions. In the second experiment based on Roget's thesaurus, on the other hand, we provided conclusive evidence that LSA is more adept at capturing particle semantics than the semantics of valence-underspecified prepositions. Taken together, these provide solid evidence that LSA produces higher-quality results in the presence of valence information. We attribute this to semantic disparities between intransitive and transitive forms of a given preposition, or to think of it in set terms, the semantics of each of the two transitivities constitutes a proper subset of that the (valence-underspecified) whole.

Due to the nature of Roget's thesaurus, we were unable to furnish evidence for the stability of transitive preposition semantics in the second experiment. The determination of alternate methods for deriving the semantics of transitive prepositions is left as an item for future research.

## 5. Related research

Past computational research on preposition semantics falls into two basic categories: large-scale symbolic accounts of preposition semantics, and disambiguation of PP sense. (Cannesson and Saint-Dizier, 2002) developed an LCS-based formal description of the semantics of 170 French prepositions in a similar vein to (Dorr, 1997), but paying particular attention to their corpus usage. (Litkowski, 2002) used digraph analysis to induce a preposition hierarchy, based upon which he proposed disambiguation rules to map preposition sense onto the hierarchy. (O'Hara and Wiebe, 2003) focused exclusively on the disambiguation task, classifying PP tokens according to their case-role in the style of the Penn treebank.

There is also a small body of computational research on prepositions in the context of verb particle constructions. Notably, (Bannard et al., 2003) used distributional similarity between VPCs and their component verbs and prepositions to predict whether the semantics of the simplex words were preserved

in VPC; indeed, the LSA similarities used herein derive from this earlier work. Similarly, (McCarthy et al., 2003) and (Baldwin et al., 2003) tested distributional similarity in various forms as a means of predicting the relative compositionality of a given VPC.

## 6. Conclusion

We have illustrated how distributional similarity methods can be used to successfully calculate inter-preposition similarity, and provided evidence for the valence-dependence of preposition semantics. More generally, we have furnished counter-evidence to the claim that prepositions are ill-suited to distributional similarity methods, in the form of the inter-preposition similarity task. Our hope is that this research will open the way to research on automatically-derived preposition thesauri to act as the catalyst in the development of preposition ontologies.

There is scope for this research to be extended in the direction of empirically-grounded evaluation of inter-preposition similarity, perhaps using human judgements. We are also interested in the impact of dependency data on the semantic classification of prepositions. These are left as items for future research.

## Acknowledgements

This research would not have been possible without the assistance of Colin Bannard, who provided the LSA similarities and provided valuable input at various points throughout the development of this research. I would also like to thank John Beavers, Aline Villavicencio and the two anonymous reviewers for their valuable input on this research. The author is supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University.

## Notes

1. By which we specifically refer to the International Corpus of English, Penn and various CLAWS tagsets.
2. <http://www.cs.utk.edu/~lsi/soft.html>
3. <http://www.umiacs.umd.edu/~bonnie/AZ-preps-English.lcs>
4. As distributed by Project Gutenberg: <http://www.gutenberg.net/etext91/roget15a.txt>
5. Interjections and phrases are also optionally listed.
6. Note that this would not be possible in WordNet, e.g., as adjectives and adverbs are listed in independent ontologies.
7. Both of which are antiquated usages.

## References

Baldwin, Timothy, Bannard, Colin, Tanaka, Takaaki, and Widdows, Dominic (2003). An empirical model of multiword expression decomposability. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Baldwin, Timothy and Villavicencio, Aline (2002). Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 98–104, Taipei, Taiwan.

Bannard, Colin, Baldwin, Timothy, and Lascarides, Alex (2003). A statistical approach to the semantics of verb-particles. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan.

Briscoe, Ted and Carroll, John (2002). Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands.

Burnard, Lou (2000). *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Cannesson, Emmanuelle and Saint-Dizier, Patrick (2002). Defining and representing preposition senses: A preliminary analysis. In *Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 25–31, Philadelphia, USA.

Deerwester, Scott, Dumais, Susan, Furnas, George, Landauer, Thomas, and Harshman, Richard (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Dorr, Bonnie J. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.

Fellbaum, Christiane, editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Extractions*. Kluwer Academic Publishers.

Harris, Zellig (1968). *Mathematical Structures of Language*. Wiley, New York, USA.

Huddleston, Rodney and Pullum, Geoffrey K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Ikehara, Satoru, Shirai, Satoshi, Yokoo, Akio, and Nakaiwa, Hiromi (1991). Toward an MT system without pre-editing – effects of new methods in **ALT-J/E-**. In *Proc. of the Third Machine Translation Summit (MT Summit III)*, pages 101–106, Washington DC, USA.

Jackendoff, Ray (1985). Semantic structure and conceptual structure. In *Semantics and Cognition*, chapter 1, pages 3–22. MIT Press, Cambridge, USA.

Kirkpatrick, Betty (1988). *Roget's Thesaurus of English Words and Phrases*. Penguin, Harmondsworth, England.

Lin, Dekang (1998a). Automatic retrieval and clustering of similar words. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, Montreal, Canada.

Lin, Dekang (1998b). Extracting collocations from text corpora. In *Proc. of the COLING-ACL'98 Workshop on Computational Terminology*, Montreal, Canada.

Litkowski, Kenneth C. (2002). Digraph analysis of dictionary preposition definitions. In *Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 9–16, Philadelphia, USA.

Lund, Kevin, Burgess, Curt, and Atchley, Ruth Ann (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–5, Pittsburgh, USA.

Mahesh, Kavi (1996). *Ontology Development for Machine Translation: Ideology and Methodology*. Technical Report MCCS-96-292, Computing Research Laboratory, NMSU.

McCarthy, Diana, Keller, Bill, and Carroll, John (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

O'Hara, Tom and Wiebe, Janyce (2003). Preposition semantic classification via Treebank and FrameNet. In *Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 79–86, Edmonton, Canada.

Pustejovsky, James (1995). *The Generative Lexicon*. MIT Press, Cambridge, USA.

Ramscar, Michael and Yarlett, Dan (2003). Semantic grounding in models of analogy: An environmental approach. *Cognitive Science*, (27):41–71.

Resnik, Philip (1995). Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

Resnik, Philip and Diab, Mona (2000). Measuring verb similarity. In *Proc. of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pages 399–404.

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Widdows, Dominic (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th*

*Annual Meeting of the NAACL (HLT-NAACL 2003)*, pages 276–83, Edmonton, Canada.

Widdows, Dominic and Dorow, Beate (2002). A graph model for unsupervised lexical acquisition. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1093–9, Taipei, Taiwan.

## Chapter 14

# A COMPUTATIONAL MODEL OF THE REFERENTIAL SEMANTICS OF PROJECTIVE PREPOSITIONS

John Kelleher

*Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI),  
Saarbrücken, Germany*

[kelleher@dfki.de](mailto:kelleher@dfki.de)

Josef van Genabith

*School of Computing,  
Dublin City University,  
Dublin 9, Ireland*

[josef@computing.dcu.ie](mailto:josef@computing.dcu.ie)

**Abstract** In this paper we present a framework for interpreting locative expressions containing the prepositions *in front of* and *behind*. These prepositions have different semantics in the viewer-centred and intrinsic frames of reference (Vandeloise, 1991). We define a model of their semantics in each frame of reference. The basis of these models is a novel parameterized continuum function that creates a 3-D spatial template. In the intrinsic frame of reference the origin used by the continuum function is assumed to be known a priori and object occlusion does not impact on the applicability rating of a point in the spatial template. In the viewer-centred frame the location of the spatial template's origin is dependent on the user's perception of the landmark at the time of the utterance and object occlusion is integrated into the model. Where there is an ambiguity with respect to the intended frame of reference, we define an algorithm for merging the spatial templates from the competing frames of reference, based on psycholinguistic observations in (Carlson-Radvansky, 1997).

**Keywords:** Frames of reference, spatial templates, potential field models, object occlusion.

## 1. Introduction

The focus of the Linguistic Interaction with Virtual Environments (LIVE) (Kelleher, 2003) project is to develop a natural language interpretive framework to underpin the development of natural language virtual reality (NLVR) systems. An NLVR system is a computer system that allows a user to interact with simulated 3-D environments through a natural language interface. People often use locative expressions to refer to objects in a visual environment. The term locative expression describes “an expression involving a locative prepositional phrase together with whatever the phrase modifies (noun, clause, etc.)” (Herskovits, 1986, pg. 7). In the simplest form of locative expression, a prepositional phrase has an adjectival role modifying a noun phrase and locates an object. Following (Langacker, 1987) we use the terms Landmark (LM) and Trajector (TR) to describe the noun phrases in a simple locative expression, see Example (1).

**Example 1** . [The book]<sub>TR</sub> on [the table]<sub>LM</sub>.

Section 2 describes the challenges in modelling projective prepositions.<sup>1</sup> Section 3 reviews previous computational work. In Section 4, we develop the LIVE model for the interpretation of projective prepositions. This model combines novel approaches to the computation of the spatial template’s origin; the gradation of a preposition’s applicability across its 3-D spatial template; object occlusion and frame of reference ambiguity resolution.

## 2. The Challenges

### 2.1 Cognitive Models of Projective Prepositions’ Spatial Templates

Psycholinguistic research indicates that “people decide whether a relation applies by fitting a spatial template to the object’s regions of acceptability for the relation in question” (Logan and Sadler, 1996, pg. 496). A spatial template is a representation of the regions of acceptability associated with a given preposition. It is centred on the landmark, and it identifies for each point in space the acceptability of the spatial relationship between the landmark and a trajector at that point. Using a spatial template, candidate trajectors can be assessed and rank-ordered by comparing the ratings of their locations in the spatial template. The candidate object whose location has the highest acceptability rating is then selected as the trajector.

Gapp’s (1995) and Logan and Sadler’s (1996) experiments reveal some of the parameters that define the constituency of a projective preposition’s spatial template. There are three areas of acceptability within a spatial template: good, acceptable and bad; the areas within a spatial template are symmetrical around

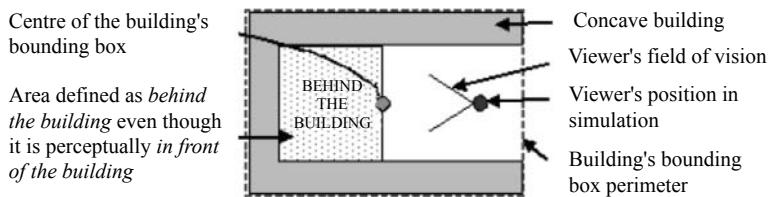


Figure 14.1. Bird's eye view of a concave building and viewer. Here, the use of the building's bounding box centre as the spatial template's origin results in an area being defined as *behind the building* even though it is perceptually *in front of the building*.

the search axis; the good and acceptable regions blend into one another; there is a sharp boundary between the acceptable and bad regions; the acceptability of a projective preposition decreases linearly as the angular deviation from the search axis increases; acceptability approaches 0 as the angular deviation approaches 90°.

In order to interpret a projective preposition in an NLVR scenario, two other factors should be integrated into the spatial template model. Firstly, the distance between each of the candidate trajectors and the landmark should be accommodated to allow the model to distinguish between candidates with the same angular deviation. Secondly, in the viewer-centred frame of reference the spatial template's origin should be located based on the user's position at the time of the utterance. The spatial template's origin is the point in space that the spatial template search axis originates from and the point from which the distances of the trajectors from the landmark are computed. Consequently, the location of the spatial template origin impacts on the acceptability ascribed to a point in the spatial template. Many previous NLVR systems (Fuhr, 2001), (Gapp, 1994), (Olivier, 1994), (Yamada, 1993) define this origin as the centroid of the landmark's bounding box.<sup>2</sup> While this approach works well for simple solid objects, applying it to more complex shapes can be problematic. For example, when applied to a concave object the centroid of the bounding box may be outside the object. This can result in paradoxical classification of regions around the landmark, see Figure 14.1.

## 2.2 Frame of Reference Ambiguity

Intrinsic to the use of a projective preposition (e.g., *in front of*, *behind*, etc.) is the definition of the direction the preposition describes. This directional constraint is referred to as the search axis. The orientation of the search axis associated with projective prepositions is dependent on the frame of reference

being used. A frame of reference consists of six half-line axes with their origin at the landmark; these axes are sometimes referred to as the base axes (Herskovits, 1986). In English, these axes are usually labelled *front*, *back*, *right*, *left*, *up* and *down*. Significantly, a frame of reference's base axes are not fixed in space, but may be rotated depending on the perspective used. Consequently, a number of frames of reference are possible. In English<sup>3</sup>, there are three different types of frames of reference: absolute, intrinsic and viewer-centred (Levelt, 1996; Levinson, 1996; Carlson-Radvansky and Irwin, 1993). Following (Levinson, 1996), we distinguish between the frames of reference based on the cardinality of their relations.

**Absolute** (extrinsic, environmental, world based) frame of reference: this is a binary reference frame that locates a trajector relative to a landmark. The labelling of the landmark's axes is dependent on salient environmental features; e.g., gravity, magnetic poles, etc.

**Intrinsic** (object-centred, landmark-based) frame of reference: involves binary relations that locate a trajector relative to a landmark. The axes of the coordinate system are oriented around the landmark based on its canonical position.

**Viewer-centred** (egocentric, relative, deictic) frame of reference: presupposes a viewpoint with ternary relations that locate an object relative to a landmark. The axes of the landmark are oriented based on a “canonical encounter” (Clark, 1973) between an observer and the landmark.

One of the difficulties for interpreting a locative expression is that many spatial expressions are common between intrinsic and viewer-centred systems. The sharing of linguistic terms across frames of reference can cause misinterpretations based on frame of reference ambiguity. Levelt (1996) uses the term coordination failure to describe such misinterpretation. In some instances, the possibility of coordination failure can be avoided by the speaker using an explicit linguistic cue. For example, the use of the determiner *the* in a noun phrase which describes a spatial region *X*, such as *the X*, implies that an intrinsic frame of reference is being used. The region denoted by *on top of X* could apply to any frame of reference described; in contrast, the region denoted by *on the top of X* could only apply to *X*'s intrinsic frame of reference (Landau and Munnich, 1998). However, explicit linguistic cues are exceptional. Consequently, if an NLVR system is going to interpret locative expressions, it must define an algorithm for handling the issue of frame of reference ambiguity.

### 3. Previous Computational Work

#### 3.1 Computational Models of Spatial Templates

If a computational model is going to accommodate the gradation of applicability across a preposition's spatial template it must define the semantics of the preposition as some sort of continuum function. A potential field model is one form of continuum measure that is widely used (Gapp, 1994; Olivier and Tsujii, 1994; Yamada, 1993). Using this approach, a model of a preposition's spatial template is constructed using a set of equations that for a given origin and point computes a value that represents the cost of accepting that point as the interpretation of the preposition. Another form of continuum model is proposed by (Mukerjee et al., 2000). In this model the continuum field is created by first defining the location of the field's global minimum. Following this, a set of concentric ellipses that use the global minimum as a fixed focus are created by varying the eccentricity of the ellipse and the position of the second focus. These concentric ellipses define the different regions of applicability within the model. Fuhr *et al.* (1998) propose a hybrid approach which uses the degree of overlap of an object with discretised regions as its measure.

Although these continuum models can distinguish between different locations within a spatial template, they are not ideal. Some of these models only work in 2-D (Mukerjee et al., 2000; Olivier and Tsujii, 1994; Yamada, 1993). (Fuhr et al., 1998) has problems distinguishing between the position of trajectors that are fully enclosed within a region. Most models (Fuhr et al., 1998; Gapp, 1994; Olivier and Tsujii, 1994; Yamada, 1993) use the centre of the landmark's bounding box as the spatial template's origin (this can lead to paradoxical interpretations, see Figure 1) and those that do not (Mukerjee et al., 2000) are dependent on locating the local minimum within the continuum field of a preposition which is problematic because the location of the local minimum varies from person to person. Furthermore, they all ignore the psycholinguistic evidence which indicates that, when frames of reference are dissociated, multiple frames of reference are activated and this multiple activation alters the constituency of the preposition's spatial template, see Section 14.4.4 (Carlson-Radvansky and Irwin, 1994) and (Carlson-Radvansky, 1997).

#### 3.2 Computational Approaches to Frame of Reference Ambiguity

In Section 2.2 we noted that if an NLVR system is going to interpret locative expressions it must define an algorithm for handling frame of reference ambiguity. In general, previous NLVR systems have adopted one of four approaches to this issue: (1) situate the discourse in domains where only simple objects with no intrinsic reference frame are modelled, e.g., the SHRDLU sys-

tem (Winograd, 1973); (2) assume a default frame of reference and force the user to adopt this for input, e.g., the Virtual Director system (Mukerjee et al., 2000) defaults to the intrinsic frame of reference if the landmark has one associated with it; (3) allow the user to switch between frames of reference if they use an explicit mark in the input, e.g., the CITYTOUR system (Andre et al., 1988); (4) assume that the frame of reference is supplied to the system *a priori*, e.g., the Situated Artificial Communicator (Fuhr et al., 1998). All of these approaches, however, either restrict the domain of the discourse or impose restrictions on the user.

## 4. The LIVE Model

In this section we describe the LIVE semantic model for the projective prepositions *in front of* and *behind*. Vandeloise (1991) observes that the prepositions *devant/derrière* are bisemic, because the relationships they describe between the trajector and the landmark in the intrinsic frame of reference are different from the ones they describe in the viewer-centred frame of reference. He defines a topological semantics for these prepositions in the intrinsic frame of reference and argues that the primary factor in the viewer-centred usages is object occlusion. While we agree with Vandeloise in his assertion that the prepositions *in front of* and *behind* are bisemic, we do not claim that object occlusion is the primary factor in the semantics of *in front of* and *behind*; rather the approach we adopt is more aligned with that of Jackendoff and Landau, who argue that while object occlusion impacts of the semantics of these prepositions, it plays “a secondary role, possibly forming a preference rule system with the directional criteria” (1992, pg. 114). Following this, we define two spatial templates for *in front of* and *behind*: one for the intrinsic frame of reference which does not consider object occlusion, and one for the viewer-centred frame of reference which does.

### 4.1 Locating the Spatial Template’s Origin

Most previous continuum models (Gapp, 1994; Olivier and Tsujii, 1994; Yamada, 1993) use the centre of the landmark’s bounding box as the spatial template origin, irrespective of which frame of reference is being used. For landmarks with complex geometries this can result in a paradoxical parsing of space (see Figure 1). In contrast with previous approaches, we define a different spatial template origin for each frame of reference.

In the intrinsic frame of reference, the spatial template origin is known to the system through *a priori* knowledge. The motivation for this is that if a person associates an intrinsic frame of reference with an object, they must have learned this intrinsic orientation based on prior experience with the object or objects of that type.

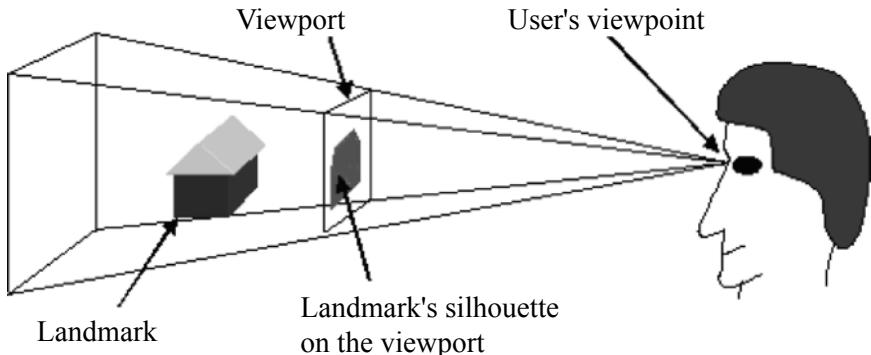


Figure 14.2. The relationships between the user's viewpoint, the viewport, a landmark's 3-D mesh and the landmark's silhouette on the viewport.

In contrast, the viewer-centred frame of reference may be applied to an object without prior knowledge of the object. From this, we argue that it is cognitively implausible to assume that a person uses a point in space whose location they do not know (i.e., the centre of the bounding box of an unfamiliar landmark) as the origin for their spatial orientation. One of the insights guiding the LIVE project is the grounding of the semantics of spatial language in visual perception. Following this, we argue that the most natural location for the origin of a projective preposition's spatial template in the viewer-centred frame or reference is the point on the landmark at the center of the landmark's silhouette as it is perceived by the user at the time of the utterance. In the terminology of 3-D graphics this point is defined as the point on the landmark's 3-D mesh that maps to the center of the landmark's silhouette on the viewport<sup>4</sup> at the time of the utterance. Figure 14.2 illustrates the relationships between the user's viewpoint, the viewport, a landmark's 3-D mesh and the landmark's silhouette on the viewport. Figure 14.3 lists the four step algorithm used to locate the point on the landmark's mesh that maps to the point at the center of its silhouette on the viewport.

The first step in the algorithm is to resolve the landmark reference. In the LIVE system, the landmark reference is resolved using the LIVE system's general algorithm for reference resolution (see (Kelleher, 2003) for details).

The second step in the algorithm is to calculate the landmark's silhouette on the viewport. We calculate the landmark's silhouette on the viewport by adapting a graphics technique called false colouring. False colouring was initially proposed by (Noser et al., 1995) as part of a navigation system for animated characters. Using a false colouring technique a system can extract information relating to the user's perception of the simulation at a given point in time.

- 1 Resolve the landmark reference.
- 2 Calculate the landmark's silhouette on the viewport.
- 3 Calculate the point at the center of the landmark's silhouette on the viewport.
- 4 Calculate the point on the landmark's 3-D mesh that maps to the center of the landmark's silhouette on the viewport.

*Figure 14.3.* The LIVE algorithm for locating the spatial template origin.



*Figure 14.4.* The image on the left is the rendered visual context. The image on the right is the false colour rendering of the landmark.

Implementing the technique involves assigning each object in the simulation a unique ID that differs from the normal colours used to render the object in the world; hence the term false colouring. An object's false colour is only used when rendering the object in the false colour rendering, and does not affect the renderings of the object seen by the user, which may be multi-coloured and fully textured. Once each object in the simulation has been assigned a false colour, whenever the system needs to examine what the user is currently seeing, a model of the user's view of the world using the false colours is rendered and the resulting image is scanned. By extracting the RGB<sup>5</sup> values found in the image, a list of objects in the image can be created. For the LIVE system we adapted and extended the false colouring technique to create a dynamic real-time model of visual salience for 3-D rendered environments; the LIVE system uses the resulting visual salience information to ground its reference resolution algorithm, see (Kelleher and van Genabith, 2004) for details. We calculate the silhouette of the landmark on the viewport by rendering the landmark by itself using its false colour (Figure 14.4 depicts the false colour silhouette of the house).

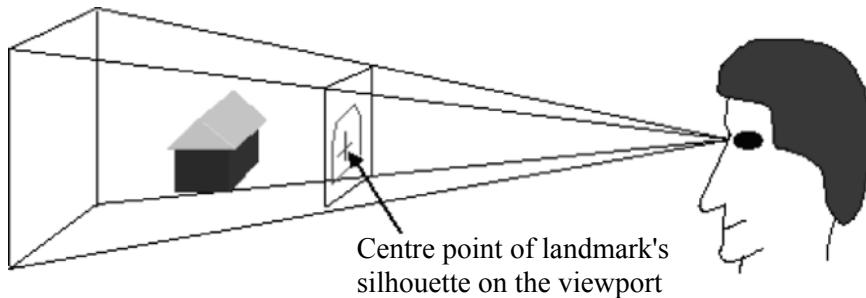


Figure 14.5. The relationships between the user's viewpoint, a landmark and the centre point of the landmark's silhouette on the viewport.

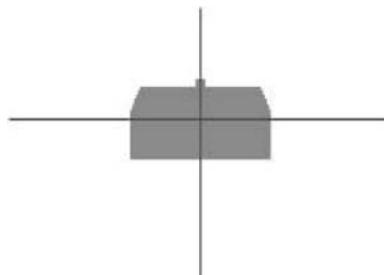


Figure 14.6. The point of intersection of the vertical and horizontal lines marks the location of the calculated center of the object's silhouette.

The next step is to calculate the coordinates of the center of the landmark's silhouette. To do this we first scan the false colour rendering of the landmark and record the maximum and minimum  $x$  and  $y$  coordinates of pixels rendered using the landmark's false colour. The coordinates of the center of the landmark's silhouette can then be calculated using Equation 14.1.

$$(14.1) \quad \text{center}(x,y) = \left( \frac{(x_{\max} - x_{\min})}{2}, \frac{(y_{\max} - y_{\min})}{2} \right)$$

Figure 14.5 illustrates the relationships between the user's viewpoint, a landmark and the centre point of the landmark's silhouette on the viewport. Figure 14.6 illustrates the point calculated as the the center of the landmark's silhouette on the viewport in our example.

The final step of the algorithm is to locate the point on the landmark at the center of its silhouette. We use a graphics technique called ray casting to locate this point. Ray casting can be functionally described as casting a ray (i.e., drawing an invisible line) from one point in a 3-D simulation in a

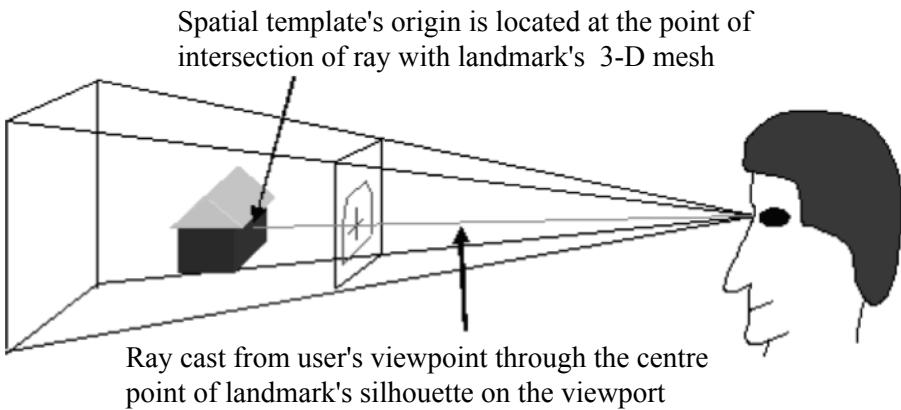


Figure 14.7. The preposition's spatial template's origin is located at the point of intersection of the ray cast from the user's viewpoint through the center point of the landmark's silhouette on the viewport and the landmark's 3-D mesh.

certain direction, and then reporting back all the intersections with 3-D object meshes and the coordinates of these intersections. To locate the point on the landmark's 3-D mesh that maps to the point at the center of its silhouette on the viewport we cast a ray from the user's viewpoint through the center point of the landmark's silhouette on the viewpoint and take the first point of intersection of this ray with the landmark's 3-D mesh as the origin of the preposition's spatial template. Figure 14.7 illustrates the casting of a ray from the user's viewpoint through the center point of the landmark's silhouette on the viewport and the intersection of this ray with the landmark's 3-D mesh. The preposition's spatial template origin is located at the point of intersection of the ray and the 3-D mesh.

## 4.2 Modelling the Gradation of a Preposition's Applicability

The two main factors that impact on the applicability of a projective preposition at a point relative to a landmark are: the angular deviation of the point from the canonical direction of the preposition's search axis and the distance of the point from the origin of the spatial template. Modelling these is further complicated by the requirement that the model should be scalable in order to accommodate different sizes of spatial configurations; e.g., the size of area described by *in front of the building* is larger than the area described by *in front of the door* (of the same building).

To model the directional constraint of a projective preposition, an algorithm for calculating the deviation of a point from a preposition's search axis must

be defined. The first stage of this process is to assign a canonical direction to each of the prepositions. We assume that the search axes for the prepositions in the intrinsic frame of reference are defined through prior knowledge. However, orienting the search axes in the viewer-centred frame of reference is dependent on the location of the user relative to the landmark at the time of the utterance. The vector originating from the spatial template's origin to the user's location describes the search axis for *in front of* in the viewer-centred frame of reference. One way of computing this vector is to convert the user's world coordinates into a set of coordinates in the local coordinate system centred on the spatial template's origin. The translated coordinates of the user's location then defines the search axis for *in front of* in the viewer-centred frame of reference. Rotating this vector by 180° gives us the search axis for *behind* in the viewer-centred frame of reference.

Having assigned a direction to each preposition, the next step in the modeling process is to devise a method for calculating the angular deviation of a candidate trajector from the search axes.  $\theta$ , the angle between two vectors  $\nu$  and  $\omega$  can be calculated using Equation 14.2:

$$(14.2) \quad \theta = \cos^{-1} \left( \frac{\nu \bullet \omega}{|\nu| |\omega|} \right)$$

where  $\nu = [x_1, x_2, x_3]$ ,  $\omega = [y_1, y_2, y_3]$ ,  $\nu \bullet \omega = (x_1 y_1, x_2 y_2, x_3 y_3)$ ,  $|\nu| = \sqrt{x_1^2 + x_2^2 + x_3^2}$ , and  $|\omega| = \sqrt{y_1^2 + y_2^2 + y_3^2}$ .

However, in order to use this equation to measure the angular deviation of a point from the search axis, the point must be converted into a vector that shares a common origin with the search axis. Applying this process to the coordinates of each of the candidate trajectors assigns each candidate an angular deviation from the preposition's canonical direction.

The distance applicability of a candidate trajector can be computed using the standard coordinate geometry distance formula for the distance between two points  $[x_1, y_1, z_1]$  and  $[x_2, y_2, z_2]$ , given in Equation 14.3:

$$(14.3) \quad \text{Dist} = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)}$$

To create the topological spatial template for a projective preposition, the angular applicability ratings must be combined with the distance applicability ratings. This is done using the algorithm listed in Figure 14.8. This algorithm requires the definition of a maximum allowable angle of deviation  $\beta$  and a maximum distance  $\gamma$ . Following the findings of (Gapp, 1995; Logan and Sadler, 1996), the maximum angle of acceptability,  $\beta$ , should be set to 90°. To date, no ratio of the maximum distance  $\gamma$  to landmark size has been identified in the research literature. We propose that  $\gamma$  be set to the distance of

**Input** A set of candidate trajectors  $\{ct_1, ct_2, \dots, ct_n\}$  each with an angular deviation  $\alpha$  and distance rating  $\delta$ ; a maximum angle of deviation for the spatial template  $\beta$ ; a maximum distance for the spatial template  $\gamma$ ; and  $\rho$  the computed scaling factor.

**Output** A set of candidate trajectors  $\{ct_1, ct_2, \dots, ct_n\}$  each with an applicability rating  $\lambda$  within the preposition's spatial template.

- 1 **let**  $\rho = 0$
- 2 **foreach**  $ct_i$ 
  - (a) **if**  $ct_i.\alpha \geq \beta$  **then**  $ct_i.\alpha = 0$  **else**  $ct_i.\alpha = 1 - (\frac{ct_i.\alpha}{\beta})$
  - (b) **if**  $ct_i.\delta \geq \gamma$  **then**  $ct_i.\delta = 0$  **else**  $ct_i.\delta = 1 - (\frac{ct_i.\delta}{\gamma})$
  - (c)  $ct_i.\lambda = ct_i.\alpha \times ct_i.\delta$
  - (d) **if**  $ct_i.\lambda > \rho$  **then**  $\rho = ct_i.\lambda$
- 3 **foreach**  $ct_i$ 
  - (a)  $ct_i.\lambda = \frac{ct_i.\lambda}{\rho}$

Figure 14.8. Algorithm for combining the angular deviation and distance scores.

the candidate trajector (simply satisfying the linguistic description of the trajector NP and within the maximum allowable angular deviation) farthest from the spatial template origin. This means that the distance from the spatial template origin does not preclude a candidate trajector from being considered as the locative expression's referent; however, it does affect its rating within the process for selecting the referent. Moreover, by allowing the spatial template's maximum distance to vary depending on the context, the spatial template is scalable to different situations. This process results in each candidate trajector being assigned a rating within the spatial template. Figure 14.9 illustrates the continuum created using the algorithm listed in Figure 14.8.

### 4.3 Perceptual Cues in the Viewer-Centred Frame of Reference

At the beginning of Section 4, we proposed that the perceptual phenomenon of object occlusion impacts on the spatial templates of the prepositions *in front of* and *behind* in the viewer-centred frame of reference. We use two rules to integrate object occlusion with the continuum model:

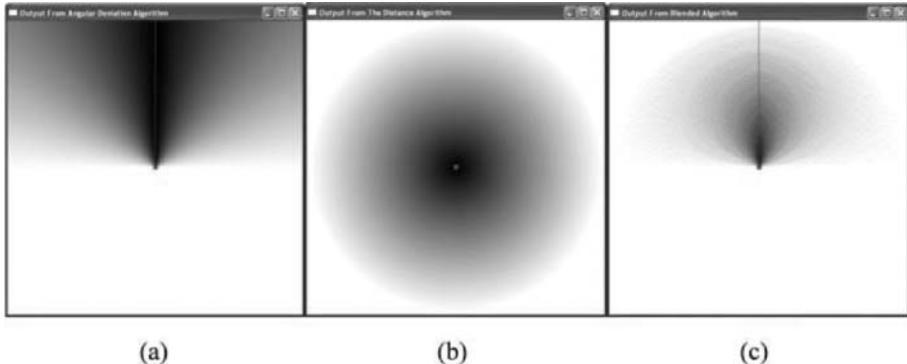


Figure 14.9. Diagrams illustrating 2-D slices of angle, distance, and amalgamated spatial template continuums. In these diagrams the darker the pixel the higher the applicability assigned to the point. The landmark is located at the centre of each image. (a) illustrates applicability gradation computed using Equation 14.2 (with search axis as the vertical axis and  $\beta = 90^\circ$ ). (b) illustrates the gradation computed using Equation 14.3 ( $\gamma$  set to a distance just inside the border of the image). (c) highlights the search axis used and illustrates the continuum created by merging the angular and distance applicabilities using the algorithm listed in Figure 14.8.

- 1 If we are interpreting a locative containing the preposition *in front of* and there is a candidate trajector which partly occludes the landmark, it is ascribed a maximum applicability rating within the viewer-centred spatial template irrespective of the rating based on the continuum model.
- 2 If we are interpreting a locative containing the preposition *behind* and there is a candidate trajector which is partly or wholly occluded by the landmark, it is ascribed a maximum applicability rating within the viewer-centred spatial template irrespective of the rating based on the continuum model.

If there is more than one candidate trajector with the maximum applicability rating we distinguish between them using a visual salience algorithm (based on size and location within the view volume), see (Kelleher and van Genabith, 2004). Moreover, if the visual salience is inconclusive (i.e., the differences in the saliences ascribed to the candidates is not sufficient to distinguish between them) we treat the locative as ambiguous and the system asks the user for clarification.

#### 4.4 Resolving Frame of Reference Ambiguity

To date there have been several sets of psycholinguistic experiments on frames of reference selection in spatial language. Carlson-Radvansky and Irwin's (1994) reports that when frames of reference are dissociated, more

than one reference frame is initially activated and these active frames compete. Carlson-Radvansky and Logan (1997) investigated the influence of frame of reference selection on the construction of a preposition's spatial template. Their findings indicate that, if there is a competition between reference frames, the construction of a preposition's spatial template in one frame of reference interferes with the construction of the spatial template in the other frame of reference. This interference between reference frames results in an amalgamated spatial template which extends over the areas covered by both of the individual spatial templates. Furthermore, the constituency of this amalgamated spatial template differs from a spatial template constructed when there is no competition: there is no good region; the acceptable regions are bigger and the bad regions are smaller; the regions that are rated as acceptable in both the viewer-centred and intrinsic frame of reference have a higher acceptability rating in the amalgamated frame of reference than those in the regions which are acceptable in only one of the individual spatial templates. Carlson-Radvansky and Logan (1997) concluded that when frames of reference are dissociated, the spatial templates constructed for each of the competing reference frames should be amalgamated using a weighting that reflects the bias towards a particular reference frame for a given preposition. With respect to the bias in this competition, Carlson-Radvansky and Irwin (1993) showed that where a preposition is canonically aligned with the vertical axis, the absolute frame of reference dominates its use, and findings in (Taylor *et al.*, 2000) indicate that, in contrast with the vertically aligned prepositions, there is a slight bias toward the intrinsic frame of reference for the horizontally aligned prepositions. Based on these psycholinguistic findings we present an algorithm (Figure 14.4) for resolving frame of reference ambiguity. Figure 14.10 illustrates the template resulting from this process.

The weighting of 2:1 towards the viewer-centred frame of reference for the vertically aligned prepositions is derived from an analysis of Carlson-Radvansky and Irwin's (1993) results. Although the work of Taylor *et al.* (2000) does not quantify the bias toward the intrinsic frame of reference for the horizontally aligned prepositions, a ratio of 1.1:1 in favour of the intrinsic frame of reference for horizontally aligned prepositions is assumed. While there is a marginal difference across this ratio, it is sufficient to prefer the intrinsic frame of reference in the event of a tie.

## 4.5 Selecting the Referent

The semantic model described in the preceding sections allows us to model the applicability of a preposition across a region. Using this model, a projective locative expression can be resolved by selecting a referent from the set of candidate trajectory based on their location within a region and object occlu-

1 **if** the frames of reference in a scene are dissociated **then**

- (a) construct a spatial template for the preposition in both frames of reference
- (b) **if** preposition = *above* or *below* **then**
  - i multiply the ratings in the viewer-centred spatial template by 2
- (c) **elseif** preposition = *in front of* or *behind* **then**
  - i multiply the ratings in the intrinsic spatial template by 1.1
- (d) assign each point an overall applicability equal to the sum of its applicability ratings in both spatial templates
- (e) select the candidate with the highest overall applicability as the referent.

frame of reference competition resolution algorithm.

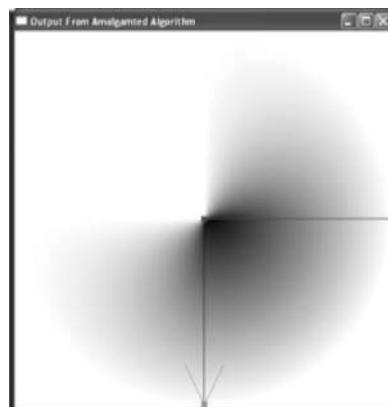


Figure 14.10. Bird's eye view of a 2-D slice of the spatial template for *in front of* created using the algorithm listed in Figure 14.4.4. The landmark is located at the centre, the viewer at the bottom, the search vector used to create the intrinsic spatial template is illustrated by the line going from the landmark to the right of the image, the search vector used to create the viewer-centred spatial template is illustrated by the line going from the landmark to the viewer's location.

sion effects. However, the abstraction used to represent the candidate trajectors impacts on this process as it affects the applicability ratings assigned to them by the model. Most previous systems have used the centre of the candidate trajector's bounding box. There are, however, problems with this abstraction for elongated objects. To account for this, we use the vertex in the candidate's 3-D mesh which has the highest applicability rating to represent each candidate. This ensures that the candidate with a point at the highest applicability will be selected as the referent.

## 5. Conclusions

In summary, the advantages of the LIVE interpretive framework described in this paper are: it avoids the problems associated with using the landmark's bounding box centre as the spatial template origin in the viewer-centred frame of reference; it offers a new model for the gradation of the preposition's applicability across a 3-D volume; it is scalable and consequently it is able to accommodate different size landmarks; it accommodates the impact of frame of reference ambiguity on the construction of a spatial template model in terms of amalgamated spatial template models and it accommodates the perceptual cue of object occlusion.

## Notes

1. For a model-theoretic analysis of locative expressions see (Zwarts and Winter, 2000).
2. An object's bounding box is the minimal rectangle that encompasses the geometry of the object.
3. Although the use of a tripartite system is common in European languages, this is not universal with many languages taking different approaches, see (Levinson, 1996) and (Levelt, 1996)
4. A viewport is the rectangular area of the display window. It can be conceptualised as a window onto the 3-D simulation.
5. RGB: Red, green and blue; the primary colours that are mixed to display the color of pixels on a computer monitor.

## References

Andre, E., Herzog, G., and Rist, T. (1988). On the simultaneous interpretation of real world image sequences and their natural language description: The system soccer. In *In Proceedings of the 8th European Conference on Artificial Intelligence (ECAI-88)*, pages 449–454. Pitmann.

Carlson-Radvansky, L.A. Logan, G. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37:411–437.

Carlson-Radvansky, L. and Irwin, D. (1993). Frames of reference in vision and language: Where is above? *Cognition*, 46:223–224.

Carlson-Radvansky, L. and Irwin, D. (1994). Reference frame activation during spatial term assignment. *Journal of Memory and Language*, 33:646–671.

Clark, H. (1973). Space, time, semantics, and the child. In Moore, T., editor, *Cognitive development and the acquisition of language*, pages 65–110. Academic Press, New York.

Fuhr, T., Socher, G., Scheering, C., and Sagerer, G. (1998). A three-dimensional spatial model for the interpretation of image data. In Olivier, P. and Gapp, K., editors, *Representation and Processing of Spatial Expressions*, pages 103–118. Lawrence Erlbaum Associates.

Gapp, K. (1994). Basic meanings of spatial relations: Computation and evaluation in 3d space. In *National Conference on Artificial Intelligence (AAAI-94)*, pages 1393–1398.

Gapp, K. (1995). Angle, distance, shape, and their relationship to projective relations. In *Proceedings of the 17th Conference of the Cognitive Science Society*.

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of prepositions in English*. Studies in Natural Language Processing. Cambridge University Press.

Jackendoff, R. and Landau, B. (1992). Spatial language and spatial cognition. In Jackendoff, R., editor, *Languages of the Mind*, pages 99–125. MIT Press.

Kelleher, J. (2003). *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. PhD thesis, Dublin City University.

Kelleher, J. and van Genabit, J. (2004). A false colouring real time visual saliency algorithm for reference resolution in simulated 3d environments. *AI Review* 21:253–267.

Landau, B. and Munnich, E. (1998). The representation of space and spatial language: Challenges for cognitive science. In Olivier, P. and Gapp, K., editors, *Representation and Processing of Spatial Expressions*, pages 262–272. Lawrence Erlbaum Associates.

Langacker, R. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites*, volume 1. Standford University Press.

Levelt, W. (1996). Perspective taking and ellipsis in spatial descriptions. In Bloom, P. and Peterson, M., Nadell, L., and Garrett, M., editors, *Language and Space*, pages 77–108. MIT Press.

Levinson, S. (1996). Frame of reference and molyneux’s question: Crosslinguistic evidence. In Bloom, P. and Peterson, M., Nadell, L., and Garrett, M., editors, *Language and Space*, pages 109–170. MIT Press.

Logan, G. and Sadler, D. (1996). A computational analysis of the apprehension of spatial relations. In Bloom, P. and Peterson, M., Nadell, L., and Garrett, M., editors, *Language and Space*, pages 493–529. MIT Press.

Mukerjee, A., Gupta, K., Nauityal, S., Mukesh, P., Singh, M., and Mishra, N. (2000). Conceptual description of visual scenes from linguistic models. *Journal of Image and Vision Computing*, 18.

Noser, H., Renault, O., Thalmann, D., and Magnenat-Thalmann, N. (1995). Navigation for digital actors based on synthetic vision, memory and learning. *Computer Graphics*, 19(1):7–9.

Olivier, P. and Tsujii, J. (1994). Quantitative perceptual representation of prepositional semantics. *Artificial Intelligence Review*, 8(147-158).

Taylor, H., Naylor, S., Faust, R., and Holcomb, P. (2000). Could you hand me those keys on the right? disentangling spatial reference frames using different methodologies. *Spatial Cognition and Computation*, 1(14):381–397.

Vandeloise, C. (1991). *Spatial Prepositions: A Case Study From French*. The University of Chicago Press.

Winograd, T. (1973). A procedural model of language understanding. In Schank, R. and Colby, K., editors, *Computer Models of Thought and Language*, pages 152–186. W. H. Freeman and Company.

Yamada, A. (1993). *Studies in Spatial Descriptions Understanding based on Geometric Constraints Satisfaction*. PhD thesis, University of Kyoto.

Zwarts, J. and Winter, Y. (2000). Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information*, 9(2):169–211.

## Chapter 15

# ONTOLOGY-BASED SEMANTICS FOR PREPOSITIONS

Per Anker Jensen

*Computational Linguistics*  
*Copenhagen Business School*  
paj.id@cbs.dk

Jørgen Fischer Nilsson

*Informatics and Mathematical Modelling*  
*Technical University of Denmark*  
jfn@imm.dtu.dk

**Abstract** This paper outlines a relation-logical compositional semantics for the meaning content of nominals using formal ontologies as semantic domains. Prepositions are conceived as denoting binary semantic role relations between concepts in the ontology. The ontology comes with ontological affinities specifying the admissible ontological combinations. The key idea is to establish a many-many relation between lexical items and nodes in an ontology. This mapping is then systematically extended to phrases, appealing to a relational compositionality principle. The paper focuses on the semantics of prepositions and prepositional phrases, examining in particular disambiguation of nominal phrases containing multiple embedded prepositional phrases, utilizing the ontological affinities. Danish, which offers a rich system of prepositions, is used in the example material.

**Keywords:** Ontological semantics, prepositional semantics, semantic roles, generative ontologies, natural language processing.

### 1. Introduction

This paper outlines a theory of the semantics of prepositions and prepositional phrases as modifiers of nouns. The semantics is based on two assumptions: on the one hand the existence of a formal ontology forming a lattice whose nodes are simple or complex concepts generated by operations such as

lattice meet, join and Peirce product. On the other hand, the existence of a finite set of universal, that is, language independent, binary role relations such as AGENT, CAUSE, PATIENT, SOURCE, TEMPORALITY and others. The role relations express possible relations among the nodes in the lattice constituting the ontology. Thereby they make possible the generation of an infinite number of ontological nodes in the lattice, thus establishing a “generative ontology”.

Each of the role relations places ontological restrictions on the arguments acting as its relata. By way of example, the complex concept *vitamin-D*[TEMPORALITY: *winter*], which represents the meaning of an English phrase like *vitamin D in the wintertime*, seems to be ontologically inadmissible according to a general principle of ontological or conceptual composition disallowing stuff-denoting concepts, e.g. vitamins, to be conceptually modified by temporal concepts. Thus *vitamin-D* [TEMPORALITY: *winter*] exemplifies an ontologically inadmissible category in violating the ontological restrictions which the TEMPORALITY relation places on its two arguments. By contrast, the complex concept *lack* [WITH RESPECT TO: *vitamin-D*, TEMPORALITY: *winter*], corresponding to an English phrase like *lack of vitamin D in the wintertime*, is sanctioned by an ontology admitting temporal modification of states.

Prepositions are conceived as each realizing a subset of the finite set of role relations. Thus, a very versatile Danish preposition like *af* ('of') realizes a subset including elements like AGENT, CAUSE, PATIENT, SOURCE, TEMPORALITY and others, whereas a preposition like *i* ('in') realizes fewer roles, among them TEMPORALITY and LOCATION.

On the theoretical basis briefly laid out above, this paper addresses, in particular, the problem of disambiguating Noun Phrases (NPs) containing one or multiple Prepositional Phrases (PPs). At the end of the paper we illustrate how our approach makes it possible on a principled basis to capture paraphrase relations among syntactically diverse constructions such as NPs containing PPs, genitive constructions, and Noun-Noun compounds.

The semantic values of NPs and PPs are expressed by nodes representing concepts in the formal ontology. PPs modifying nouns are treated as expressing conceptual specializations of the concept expressed by the noun. The potentially infinite syntactic complexity of NPs containing PPs calls for a compositional semantics reflecting this infinity. This is achieved by generating an infinity of nodes in the ontology as indicated above. The discarding or (partial) disambiguation of NPs containing PPs is carried out by performing a check on the ontological admissibility of the type of the arguments of a role relation. For present purposes metonymy and live metaphors are disregarded.

The perspectives of this research are twofold: In the short term, to improve information extraction and retrieval by enabling the system to generate and compare relatively fine-grained ontological descriptions for queries and text items in databases or knowledge bases, cf. (Andreasen *et al.* 2002). The

long-term objective is to provide a contribution to a formal, ontology-based semantics for a more comprehensive fragment of natural language expressions.

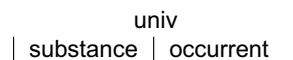
The structure of the remainder of this paper is as follows: Section 2 briefly elaborates the notion of 'formal ontology' by an example; section 3 addresses the question of how to establish a systematic relationship between natural language items and the concepts of a proposed ontology; section 4 discusses formal meaning ascription to phrases using principles of relational composition. Section 5 elaborates on the notion of generative ontology involving feature structures. Section 6 introduces the notion of ontological affinities, and 7 sets up an ontological semantics for nominals. Section 8 practices the semantics on a selection of Danish NPs containing modifying PPs. Section 9 illustrates how paraphrases among syntactically diverse construction types may be recognised by the approach proposed. Section 10 sums up the results of the paper.

## 2. Formal ontologies

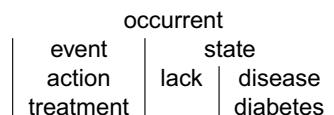
In the present context an ontology is conceived of as a general description of the concepts in a domain structured by an inclusion ordering of the concepts in terms of sub- and superconcepts, see further e.g. (Guarino 1995) and (Smith 2002). Ontologies are abstract classifications not subject to physical requirements on linear ordering and arrangement. This facilitates making integrated use of more flexible and complex non-hierarchical forms of categorization such as lattices, where a concept may have more than one immediate superconcept. For instance, in the ontology shown below, **object** and **stuff** may overlap in a category of **portion**. In the ontology, closely related concepts are placed closer to each other in the lattice than non-related concepts, and we use nominalised forms consistently in order to support the relation of concept inclusion.

## 2.1 Skeleton ontology

By way of illustration, below we have sketched a fragment of a so-called skeleton ontology, that is to say, only the inclusion relation (the *isa*-relation) is considered. As it appears, the universal top category divides into material concepts, substance, and occurrent.



The category **occurrent**, in turn, divides as follows



meaning that the category **diabetes** is a subcategory of **disease**, which is a **state**, etc.

The category **substance** has the subcategories **object** and **stuff** with their subcategories

person	object	portion	portion	stuff	foodstuff
child	organ	liver1	medicine	vitamin	liver2
				tocopherol	

Here it should be observed that **portion** appears twice by virtue of its possessing a dual nature of object and stuff, and implying that the skeleton ontology is not strictly hierarchical.

Ontologies in our conception are extra-linguistic, language independent logical structures. Thus the above labels **object**, **stuff**, etc., are names of the meta-language. The factual relationship that any  $p$  is a  $q$ , corresponds to the logical clause  $q(X) \leftarrow p(X)$ .

However, in the subsequent formalization object-language concepts  $p$ ,  $q$ , etc., appear encoded as terms in a logic with distinguished ontological predicates as in  $isa(p, q)$ . This meta-logical set-up complies with definite clause grammars.

## 2.2     Ontology with semantic roles

In addition to the relation of concept inclusion, the ontology comprises a number of binary role relations holding between pairs of concepts, e.g. relations such as **AGENT**, **PART-OF**, **PATIENT**, **LOCATION**. In our formalization these roles are represented as attributes attached to the concept labels as explained in section 5. This enrichment facilitates the adoption of ontologies as semantic target domains.

## 3.     The relation between lexicon and ontology

The relationship between a natural language lexicon and an ontology is established by a relation *lex*. In the case of nouns the *lex* relation is realized as:

*lex(noun, concept)*

where *noun* is an appropriately normalized (lemmatized) word form and *concept* is a node label in the ontology. Thus we may have examples like *lever* (*liver*), represented as:

*lex(lever, liver 1)  
lex(lever, liver 2)*

displaying the polysemy of this word between its organ and foodstuff senses, cf. the ontology outlined above. Conversely, synonymy can be expressed thus:

$\text{lex}(E\text{-vitamin}, \text{tocopherol})$   
 $\text{lex}(\text{tokeferol}, \text{tocopherol})$

The relation  $\text{lex}$  establishes a multi-valued mapping of nouns into nodes in the proposed ontology forming the semantic domain in our semantics, and, conversely,  $\text{lex}$  establishes a multi-valued relation from the concepts in the ontology to the lexical items. The mapping from the Danish lexical items *lever*, *E-vitamin*, and *tokeferol* to the appropriate nodes in the ontology can be illustrated like this:

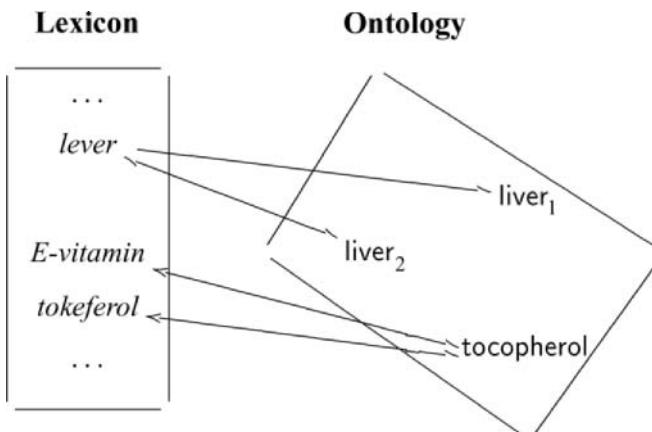


Figure 15.1. Lexicon and Ontology

Thus  $\text{lex}$  is not a genuine lexical relationship in the sense of holding between lexical items, cf. e.g. (Cruse, 1986), rather, it is a relation between lexical items and non-linguistic objects, namely concepts which are abstractions of real world entities. Thus, our approach appeals to a realist account of language.

In the case of prepositions, the  $\text{lex}$ -relation is somewhat different from what we have just seen for nouns. We propose that the semantics of prepositions should be expressed as subsets of a presumed universal set of role relations holding between the concepts in the ontology. As already mentioned, among the role relations we find such as the PATIENT relation, occurring in examples like *behandling af sukkersyge* ('treatment of diabetes'), where diabetes assumes the role of PATIENT, which is but one out of the many relations which can be expressed by the preposition *af* ('of'). The representation in this case would allow the second argument of the  $\text{lex}$ -relation to be the name of a relation:

$$\begin{aligned}lex(af, pnt) \\ lex(af, agt) \\ lex(af, pof)\end{aligned}$$

## 4. Formal meaning ascription

Generally speaking, the process of mathematical (logico-algebraic) meaning ascription to a class of linguistic phrases may be achieved by a binary semantic relation holding between the pertinent phrases and appropriate meaning-carrying mathematical objects:

$$sem(phrase, semantic\_object)$$

In traditional Montagovian semantics, cf. e.g. (Dölling, 1995), these objects take the form of set- or function constructs over abstract domains of primitive entities. In the present exploration, the semantic objects are constituted by algebraic terms in an algebraic lattice enriched by feature structures. This means that the semantic objects may be conceived as nodes of a formal ontology, which becomes generative by virtue of the recursive productivity of terms. The simplest case of an ontological semantics for nouns is trivially formalized by appealing to the mapping relation *lex*, which holds between lexical items and corresponding nodes in the ontology. This is expressed as the definite logical clause

$$sem(noun(N), C) \leftarrow lex(N, C)$$

with variables identified as upper case letters being universally quantified by default. In the more general semantics developed in section 7 this clause covers the case of an NP realized by a single noun.

### 4.1 Relational compositionality

A fundamental principle in phrase-directed formal semantics is the notion of compositionality, according to which the object representing the meaning of a phrase is formed systematically by (functional) composition of the meanings of its sub-phrases, cf. e.g. (Janssen, 1997). We employ relational meaning composition, generalizing the notion of functional composition. Relational composition facilitates the handling of ambiguities inherent in the phrase or stemming from the absence of an embracing context for the phrase. Disambiguation and elimination of compositionally ill-formed formal meaning objects, is then to take place in a constraint resolution process as explained in section 7 below. Therefore, in general the *sem*-relation mentioned above associates multiple nodes with a given phrase in the case of homonymous and polysemous expressions. The multiplicity of readings is reduced when phrases are combined with other phrases in the context, subject to the category con-

straints expressed in the ontology. Lexical items and phrases expressing category mistakes lack associated nodes, and, conversely, in the ontology there are concept nodes without any associated lexical manifestation. The remainder of this paper explores the ontology-based semantic relationship between natural language phrases and their conceptual content, and presents formalizations for an interesting fragment of nominal constructions.

## 5. Generative ontologies with feature structures

A skeleton ontology comprises a finite number of concepts, but linguistic phrases call for a semantic target domain comprising an infinitude of compound concepts. The process of forming compound concept terms may be achieved by introduction of ontological relationships acting as binders between the concepts in the skeleton ontology. The compound algebraic terms introduced in the ontology may then be shaped as feature structures, cf. (Carpenter, 1992) and (Rounds, 1997):

$$c \left[ \begin{array}{l} r_1 : \varphi_1 \\ \dots \\ r_m : \varphi_m \end{array} \right]$$

where the attribute names  $r_i$  are semantic roles holding between the header concept  $c$  and the  $c$ -concept sub-terms  $\varphi_i$ . Argument concept terms are themselves either concept constants (non-compound concepts in the skeleton ontology) or compound concept terms. Thus, the feature structure may be nested, giving rise to an infinity of ever more specialized concepts and thus endowing the ontology with generativity. Attachment of a feature structure to a concept formally represents a restriction on that concept, so that the compound term attaches to a node below the header concept in the ontology:

$$\text{isa}(C[r : C'], C)$$

The resulting feature-structured semantic objects are reminiscent of the semantic structures in (Davidson, 1967) as well as in the generative lexicon approach (Pustejovsky, 1995). However, the semantic terms come here as an integral part of a supporting formal ontological framework. In the logico-algebraic context of distributive lattices the feature structure  $c[r_1 : \varphi_1]$  may be formalized as an algebraic conjunction (lattice meet):

$$c \wedge (r_1 : \varphi_1)$$

where the infix operator ( : ) is the so-called Peirce product, see (Brink *et al.* 1994), with  $r_1$  being a binary relation, cf. also the ONTOLOG proposal of (Nilsson, 2000). Thus the adopted semantic domain set-up is logico-algebraic in the tradition of, e.g., (Bach, 1986) and (Link, 1992). Actually, this form of

relational algebra is an algebraic counterpart of basic description logic. It is important to realize that the feature names  $r$  function also as binary relations between concepts in the ontology spanned by the terms.

## 6. Ontological affinities and generative ontologies

Formally, complex concept terms may be formed by freely combining concept labels with roles into feature terms. However, for ontological reasons, semantic roles do not combine freely with concepts in the ontology. For instance, temporal concepts apply to events but not to substances. The ontologically admissible combinations of concepts and role relations are to reflect preclusion of category mistakes, cf. e.g. (Ryle, 1949) and (Sommers, 1963). They may be declared as so-called affinities imposed on the ontology:

$$\text{affinity}(c, r, c')$$

This affinity licenses the compound term  $c[r : c']$  as an admissible concept through the rule:

$$\begin{aligned} \text{wfconcept}(D[R : D']) \leftarrow \\ \text{affinity}(C, R, C') \wedge \text{isa}(D, C) \wedge \text{isa}(D', C') \end{aligned}$$

which establishes an inheritance principle for affinities, assuming monotonic inheritance. This means that a stated affinity subsumes all its ontological specialisations. We further appeal to an ontological well-formedness principle admitting simultaneous presence of distinct roles:

$$\begin{aligned} \text{wfconcept}(C[R_1 : C_1, R_2 : C_2]) \leftarrow \\ \text{wfconcept}(C[R_1 : C_1]) \wedge \\ \text{wfconcept}(C[R_2 : C_2]) \wedge \\ \text{distinct}(R_1, R_2) \end{aligned}$$

For instance, the complex concept *lack* [WITH RESPECT TO: vitamin-D, TEMPORALITY: winter] is licensed by this clause in that the roles WITH RESPECT TO and TEMPORALITY are distinct. Some roles, however, may appear repeatedly at the same level in a feature structure. Such cases would have to be accommodated by additional clauses. Furthermore, for all role relations an “inversion principle” is in effect such that for any affinity  $\text{affinity}(c, r, c')$  it is complemented by  $\text{affinity}(c', r^{-1}, c)$ , where  $r^{-1}$  is the inverse relation of  $r$ . The inverse relation is bound to exist mathematically. Sometimes it has its own role name as in the pairs CAUSE/CAUSED-BY, COMPRIZE/PART-OF, etc., in other cases it does not.

As mentioned in section 5, a generative ontology comes about by extending the skeleton ontology with affinities licensing the ontologically admissible complex concepts in the top ontology. Consider a top ontology with ontologi-

cal principles to the effect that substances are ontologically admissible parts of substances, events can ontologically admissibly comprise events (the inverse of the POF relation), and an object (which is a subtype of substance, cf. the ontology outlined in section 2) can play the role of agent in an event. Formally such affinities are stated as factual clauses like:

<i>affinity</i> (substance, <i>pof</i> , substance)	(PART-OF)
<i>affinity</i> (event, <i>pof</i> , event)	(PART-OF)
<i>affinity</i> (event, <i>cmp</i> , event)	(HAVING-PART)
<i>affinity</i> (event, <i>agt</i> , substance)	(AGENT)
<i>affinity</i> (state, <i>tmp</i> , time)	(TEMPORALITY)
<i>affinity</i> (lack, <i>wrt</i> , any)	(WITH RESPECT TO)

The ontology becomes potentially infinite due to recursive definitions (as for POF above) similar to recursive production rules for formal languages. These affinities together with the *isa*-spanned skeleton ontology restrict the free generation of compound concept terms as specified by the predicate *wfconcept*. Methodologically, the empirical study of admissible affinities (cf. category mistakes) may thus guide the design of top ontologies.

Alternatively, the ontology may be explicated and specified as a grammar as in (Andreasen & Fischer Nilsson, 2003), cf. also (Jackendoff, 1990). If so, the concepts become non-terminals, and the syntactic derivation relation forms the opposite of the *isa*-relationship.

The analysis of French instrumental prepositions in (Mari, 2004) and (Mari & Saint-Dizier 2004) aims at accounting for the distribution of instrumental prepositions and the constraints they impose on their environment. Their language-dependent, bottom-up approach contrasts with our language-independent, ontology-based attempt at a top down analysis. The former approach distils out details in the interaction of senses, whereas the present methodology aims at an abstract language independent model for compatibility of concepts. The two approaches differ in their formalisation principles (functional vs. relational) making it difficult to assess their compatibility.

## 7. Compositional ontological semantics for nominals

The meaning of a syntactically and ontologically well-formed NP is represented as one or more nodes in the proposed ontology. In principle, of course, the aim is for the generative ontology to account for all types of NPs, e.g. with Noun-Noun compounds as heads, adjectival modifiers, genitival determiners and all combinations of these. Here we focus on nominal phrases with embedded PPs. In contrast to the Montagovian tradition, which emphasizes the logical function of NPs in the sentence, we focus on the conceptual content of each NP, which is why, under our view, there is no need to pay special at-

tention to the semantic contribution of determiners, contrast e.g. (Francez & Steedman, 2004).

In what follows we treat NPs with optional PPs, limited to two for the sake of simplicity and without essential loss of generality. We leave out of account any determiners the NPs might contain. As mentioned in section 3, the relation *lex* accounts for the relationships between prepositions and semantic roles forming a subontology. This relation is many-many since a preposition may realize different roles in different contexts, and the same role may be realized by different prepositions. Logical clauses for a compositional onto-semantics:

$$\begin{aligned}
 & \text{sem}(\text{noun}(N), C) \leftarrow \text{lex}(N, C) \\
 & \text{sem}(n\_pp(N, [\text{Prep}, \text{Np}]), C[R : C_1]) \leftarrow \\
 & \quad \text{lex}(N, C) \wedge \text{lex}(\text{Prep}, R) \wedge \\
 & \quad \text{sem}(\text{Np}, C_1) \wedge \\
 & \quad \text{wfconcept}(C[R : C_1]) \\
 & \text{sem}(n\_pp\_pp(N, [\text{Prep}_1, \text{Np}_1], [\text{Prep}_2, \text{Np}_2]), \\
 & \quad C[R_1 : C_1, R_2 : C_2]) \leftarrow \\
 & \quad \text{lex}(N, C) \wedge \text{lex}(\text{Prep}_1, R_1) \wedge \text{lex}(\text{Prep}_2, R_2) \wedge \\
 & \quad \text{sem}(\text{Np}_1, C_1) \wedge \text{sem}(\text{Np}_2, C_2) \wedge \\
 & \quad \text{wfconcept}(C[R_1 : C_1, R_2 : C_2])
 \end{aligned}$$

Recall that an NP containing two PPs is inherently syntactically ambiguous, since the PPs may or may not be nested. Thus both the second and the third clause potentially apply to the case of two PPs as exemplified in the next section. Here *wfconcept* functions as an admissibility condition: in the logical constraint resolution the affinity check assists lexical and structural disambiguation.

## 8. Prepositions and semantic roles in Danish

Affinity declarations can be viewed as admissibility conditions on semantic role relations in the sense that two concepts can only be related by a role relation provided that each concept is of a type or subtype licensed by an affinity declaration.

### 8.1 Language independent semantic roles

In the present framework, the semantics of prepositions is modeled by algebraic sums of roles picked from a presumed finite, universal set. The exact membership and nature of this set is a matter of dispute, cf. (Wechsler 1995: Chapter 1). The following table is intended to give an impression of some of those role relations which we consider plausible candidates for membership of

the universal set (cf. (Fillmore, 1968: 24-25), (Stockwell *et al.*, 1973: Chapter 2), (Somers, 1987), (Sparck Jones & Boguraev, 1987), (Madsen *et al.*, 2001)):

Role-relation	Abbreviation	Description
AGENT	AGT	Animate being acting intentionally
CAUSE	CAU	Inanimate force/actor
CAUSED-BY	CBY	Inverse CAU
PATIENT	PNT	Affected entity. Effected entity
PART-OF	POF	Part of whole. Member of set
COMPRIZE	CMP	Inverse POF. Whole constituted of parts
BY MEANS OF	BMO	Means to end. Instrument
SOURCE	SRC	Source. Origin. Point of departure
PURPOSE	PRP	Purpose
LOCATION	LOC	Place. Position
TEMPORALITY	TMP	Temporal anchoring. Duration. Inception etc.
MATERIAL	MAT	Material
CHARACTERIZE	CHR	Property ascription

The relations mentioned here should be regarded as forming a top ontology for role relations, while not mentioning their possible sub-roles. For instance, the TMP role is a super-role of INCEPTION, CULMINATION, DURATION and others. Likewise, CHR serves as cover term for a number of sub-roles expressible, for instance, by adjectives denoting colour, size, disposition, etc. In sum, the table above presents only a crude approximation to the very refined system of universal, language independent relations, and we currently make the simplifying assumption that the relations mentioned here form their own “flat” lattice, which in further research will be refined to form a more complex structure of relations.

## 8.2 Danish prepositions

Turning next to the description of the semantics of a subset of Danish prepositions, the table below indicates some of the role relations which each preposition may express<sup>1</sup>.

Preposition	RoleSet	Example	Gloss
<i>af</i>	AGT	<i>Behandling af læge</i>	Treatment <u>by</u> physician
	PNT	<i>Behandling af børn</i>	Treatment <u>of</u> children
	POF	<i>Siden af hovedet</i>	The side <u>of</u> the head
	MAT	<i>Pude af læder</i>	Cushion <u>of</u> leather
<i>i</i>	LOC	<i>Betændelse i øjnene</i>	Inflammation <u>of</u> the eyes
	TMP	<i>I to dage</i>	For two days
	POF	<i>Celler i øjet</i>	Cells <u>in</u> the eye
<i>med</i>	BMO	<i>Behandling med medicin</i>	Treatment <u>with</u> medicine
	CHR	<i>Børn med diabetes</i>	Children <u>with</u> diabetes
<i>fra</i>	SRC	<i>Blødning fra tarmen</i>	Haemorrhage <u>from</u> the intestine
	TMP	<i>Fra sidste år</i>	From last year
	POF	<i>En person fra staben</i>	A person <u>from</u> the staff

By means of the *lex*-relation introduced in section 3, we can now posit lexical entries for nouns

*lex(behandling, treatment)*  
*lex(brn, children)*  
*lex(sukker syge, diabetes)*  
*lex(medicin, medicine)*

and for the prepositions *af, med, i, fra*

*lex(af, agt), lex(af, pnt), lex(af, pof), lex(af, mat)*  
*lex(med, bmo), lex(med, chr)*  
*lex(i, loc), lex(i, tmp), lex(i, pos)*  
*lex(fra, src), lex(fra, tmp), lex(fra, pof)*

Applying the rules of composition given in section 7, ontology-based semantic representations for N-PP structures (with possible embedded PPs in the PP following the head noun) and for structures with consecutive non-embedded N-PP-PP structures can now be generated.

### 8.3 The case of one PP

Consider first the example *behandling af børn* (treatment of children). The *n\_pp* rule will return two ontologically admissible results. First we get **treatment** [AGT: **children**]. This representation is licensed since children are ontologically admissible agents. They may, for instance, minister treatment to their pets, and so the relevant affinity declaration for the AGT-relation, i.e. *affinity(event, agt, intentional\_agent)* allows this. The second admissible conceptual representation is: **treatment** [PNT: **children**]. Generally speaking, there are very few restrictions on what is an ontologically admissible patient. Anything from occurrents (comprising events and states) to physical objects seems to be a possible patient: One can treat sorrows as well as sores. For present purposes, we shall allow anything (**univ**) to be a patient, assuming the affinity declaration: *affinity(event, pnt, univ)*. On the other hand, the following representations will be rejected, since no affinity will allow them. First, **treatment** [POF: **children**] is ontologically ill-formed, since physical objects like children cannot form part of events. Quite similarly, **treatment** [MAT: **children**] is ruled out since an event such as a treatment is not made of physical objects like children.

### 8.4 The case of multiple PPs

We shall not consider the following two complex examples in detail, but restrict ourselves to mentioning that the *n\_pp* and *n\_pp\_pp* clauses together

allow conceptual representations with non-nested as well as nested role relations and combinations of these, too. Consider the phrase *behandling af børn med medicin* (treatment of children with medicine). Among other results, the *n\_pp\_pp* clause will return the following ontologically admissible representation in which the relations expressed by the prepositions *af* and *med*, respectively, are both seen to pertain to treatment, that is, they are not nested. A rough paraphrase of the representation below is *treatment affecting children and carried out by means of medicine*:

treatment [PNT: children, BMO: medicine]

This should be compared with an example like *behandling af børn med sukkersyge* (treatment of children with diabetes). In this case, consecutive applications of the *n\_pp* rule will return the following nested conceptual structure. Paraphrasing the conceptual structure, we get *treatment affecting children having the diabetes-property*:

treatment [PNT: children [CHR: diabetes]]

In the case of *behandling af børn med sukkersyge* (treatment of children with diabetes), the *n\_pp\_pp* clause would return:

treatment [PNT: children, CHR: diabetes]

were it not for the fact that a state of illness cannot be a property of a treatment, that is, there is no affinity declaration combining process, CHR, and state licensing the concept

treatment [CHR: diabetes].

Finally, consider *behandling med medicin af børn med sukkersyge* (treatment with medicine of children with diabetes). In this case the clauses would interact to return the ontologically admissible representation:

treatment [BMO: medicine, [PNT: children [CHR: diabetes]]]

where we have a nested compound concept at the same level as a non-nested one, yielding the interpretation: *treatment (carried out) by means of medicine and affecting a child characterized by the diabetes-property*.

## 9. Identifying paraphrases

The general principle of compositional semantics appealed to in the present approach can be expressed as the clause

$$\begin{aligned}
 \text{sem}(\text{phrase}(P_1, P_2), C) \leftarrow \\
 \text{sem}(P_1, C_1) \wedge \text{sem}(P_2, C_2) \wedge \\
 \text{combine}(C_1, C_2, C)
 \end{aligned}$$

where  $P_1$  and  $P_2$  are subphrases and the resulting semantic concept value  $C$  is formed by  $\text{combine}(C_1, C_2, C)$ , which in turn appeals to well-formedness check as explicated in section 6.

It is our contention that this general principle subsumes genitive constructs and Noun-Noun compounds in addition to the N-PP constructions already discussed. Consider the examples *sygdommens behandling* (lit. the disease's treatment) and *sygdomsbehandling* (lit. disease treatment), which are near-paraphrases of N-PP construction *behandling of sygdommen* (treatment of the disease). Since the three NPs are paraphrases, they should be assigned the same conceptual representation, which, according to the compositional principles given in section 7, is **treatment[PNT: disease]**.

For the case of the genitive construction<sup>2</sup> this is achieved by the clause

```
sem(np_gen(N1, N2), C2[R : C1]) ←
  lex(N1, C1) ∧ lex(N2, C2) ∧
  genitive_abduce(C1, C2, R)
```

where *genitive\_abduce* suggests possible role relations yielding, among others, **treatment[PNT: disease]** and **treatment[POF: disease]** as ontologically admissible, although it is disputable whether the latter is ontologically acceptable. This appeal to abductive derivation of a role should be contrasted with the use of *lex* in section 7 for providing the role relation associated with the pertinent preposition. We envisage a similar abductive treatment of Noun-Noun compounds.

## 10. Conclusion

We have presented a theory of the semantics of prepositions based on the notion of generative ontologies. Generative ontologies are formed by means of rules for combining concepts by means of binary semantic roles. The rules given as clauses express ontological affinities which rule out category mistakes. In our ontological semantics the semantic objects are formal, logico-algebraic objects in contrast to the informal image schemata of cognitive semantics in (Langacker, 1987) and (Talmy, 2000). The notion of generative ontology is inspired by the generative grammar paradigm and provides semantic domains for a compositional ontological semantics for NPs containing PPs. On the other hand, in contrast to traditional logical semantics, which strongly emphasizes the semantic contribution of determiners, our ontological semantics places decisive weight on the conceptual semantics of the nominal parts of NPs and their modifiers such as PPs. We have outlined how, in principle, the present approach can be extended largely unchanged to genitive constructions and Noun-Noun compounds.

## Acknowledgement

This work is part of the ONTO QUERY project (OntoQuery, 2004) supported by a grant from the Danish National Science Boards.

## Notes

1. (Diderichsen, 1971: 70) enumerates the following prepositions as the central ones in Danish: *ad* ('along'), *af* ('of'), *efter* ('after'), *for* ('for', 'in front of'), *fra* ('from'), *gennem* ('through'), *hos* ('at'), *i* ('in'), *med* ('with'), *mellem* ('between', 'among'), *mod* ('against'), *om* ('about'), *over* ('over'), *på* ('on'), *til* ('to'), *under* ('under'), *ved* ('near', 'at') . Others are: *før* ('before'), *imod* ('against'), *nær* ('near'), *omkring* ('around'), *uden* ('without').

2. For a comprehensive, formal account of prenominal genitives see (Vikner & Jensen, 2002).

## References

Andreasen, T., Jensen, P. Anker, Nilsson, J. Fischer, Paggio, P., Pedersen, B. Sandford & Thomsen, H. Erdman, (2002), "Ontological Extraction of Content for Text Querying." In Andersson, B., Bergholtz, M. & Johannesson, P. (eds.), *Natural Language Processing and Information Systems, LNCS 2553*.

Andreasen, T. & Nilsson, J. Fischer (2003), "Grammatical Specification of Domain Ontologies." Forthcoming in *Data & Knowledge Engineering*.

Bach, E. & Harms, R. (1970), *Universals in Linguistic Theory*. Holt, Rinehart and Winston. London. New York. Sydney. Toronto.

Brink, C., Britz, K. & Schmidt, R.A. (1994), "Peirce Algebras." *Formal Aspects of Computing*, Vol. 6. pp. 339-358.

Carpenter, B. (1992), *The Logic of Typed Feature Structures*. Cambridge University Press.

Cruse, D.A. (1986), *Lexical Semantics*. Cambridge University Press.

Davidson, D. (1967), "The Logical Form of Action Sentences". In: (Rescher, 1967).

Diderichsen, P. (1971), *Elementær dansk grammatik*. Gyldendal. København.

Dölling, J. (1995), "Ontological domains, semantic sorts and systematic ambiguity." *International Journal of Human-Computer Studies*, 43, pp. 785-807.

Fillmore, C. (1968): "The Case for Case". In (Bach & Harms 1970). pp. 1-88.

Francez, N. & Steedman, M. (2004), Cascaded Contextual Preposition Phrases, Present volume.

Guarino, N. (1995), "Formal Ontology, Conceptual Analysis and Knowledge Representation", *Int. J. of Human-Computer Studies*, 43(5/6), pp. 907-928.

Jackendoff, R. (1990), *Semantic Structures*. MIT Press.

Janssen, Th. (with an appendix by B. H. Partee) (1997), "Compositionality." In (van Benthem et al. 1997), pp. 417-473.

Jensen, Per Anker, Nilsson, J. Fischer & Vikner, C. (2001), "Towards an Ontology-based Interpretation of NPs." In (Jensen & Skadhauge 2001), pp. 43-55.

Jensen, P. Anker & Skadhauge, P. (eds.) (2001), *Ontology-Based Interpretation of Noun Phrases. Proceedings of the First International OntoQuery Workshop*. Department of Business Communication and Information Science. University of Southern Denmark, Kolding.

Langacker, R. (1987), *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*, Stanford U.P.

Link, G. (1998), *Algebraic Semantics in Language and Philosophy*. Center for the Study of Language and Information, Stanford.

Madsen, B. Nistrup, Pedersen, B. Sandford, Thomsen, H. Erdman (2001), “Semantic Relations for OntoQuery.” In (Jensen & Skadhauge, 2001), pp. 57-88.

Mari, A. (2004), What do “Instrumentality” and “Manner” have in Common ?, present volume.

Mari, A. & Saint-Dizier, P. (2004), A Conceptual Semantics for Prepositions denoting Instrumentality, present volume.

Nilsson, J. Fischer: “A Logico-Algebraic Framework for Ontologies, ONTOLOG.” In (Jensen & Skadhauge 2001), pp. 11-38.

OntoQuery project net site. <http://www.ontoquery.dk>.

Pustejovsky, J. (1995), *The Generative Lexicon*. MIT Press.

Rescher, N. (ed.) (1967), *The Logic of Decision and Action*. Pittsburgh U.P.

Rounds, W.C. (1997), “Feature Logics.” In (van Benthem & Meulen, 1997), pp. 477-533.

Ryle, G. (1949), *The Concept of Mind*, Penguin, 1973.

Smith, B. (2002), “Ontology and Information Systems” <http://ontology.buffalo.edu/smith/>.

Somers, H. L. (1987), *Valency and Case in Computational Linguistics*, Edinburgh University Press.

Sommers, F. (1963), “Types and Ontology.” *Philosophical Review*, 72, pp. 327-363.

Sparck Jones, K. & Boguraev, B. (1987), A Note on a Study of Cases, *Computational Linguistics*, Vol. 13, Numbers 1-2, pp. 65-68.

Stockwell, R., Schachter, P. & Partee, B.H. (1973), *The Major Syntactic Structures of English*. Holt, Rinehart and Winston, Inc. New York. Chicago. San Francisco. Atlanta. Dallas. Montreal. Toronto. London. Sydney.

Talmy, L. (2000), *Toward a Cognitive Semantics*, Vol. 1-2, MIT Press.

van Benthem, J. & ter Meulen, A. (eds.) (1997), *Handbook of Logic and Language*, Elsevier.

Vikner, C. & Jensen, P. Anker (2002), A Semantic Analysis of the English Genitive. *Interaction of Lexical and Formal Semantics*, *Studia Linguistica*, 56(2). pp. 191-226.

Wechsler, S. (1995), *The Semantic Basis of Argument Structure*. CSLI Publications.

# Chapter 16

## ANALYSIS AND INTERPRETATION OF THE JAPANESE POSTPOSITION *NO*

Ryusuke Kikuchi

*Graduate School of Computer and Cognitive Sciences, Chukyo University  
101 Tokodachi, Kaizu-cho, Toyota-shi, Aichi-ken 470-0393, Japan*

[kikuchi@cyber.sccs.chukyo-u.ac.jp](mailto:kikuchi@cyber.sccs.chukyo-u.ac.jp)

Hidetosi Sirai

[sirai@sccs.chukyo-u.ac.jp](mailto:sirai@sccs.chukyo-u.ac.jp)

**Abstract** The Japanese postposition *no* is one of the most frequently used postpositions in Japanese. The meaning of this postposition corresponds roughly to the preposition “of” and the possessive marker “s” in English. Typically, *no* forms a noun phrase,  $NP_1 \ no \ NP_2$ . The meaning of this noun phrase depends not only on the semantic properties of *NPs* occurring in the construction but also on contextual information.

In this paper, we conduct a syntactic and semantic analysis of *no*. We claim that there are two kinds of *no* in  $NP_1 \ no \ NP_2$  constructions: one is a marker of complement, and the other is the head of an adnominal phrase,  $NP_1 \ no$ , which modifies a noun phrase  $NP_2$ . From a semantic point of view, the meaning of an  $NP_1 \ no \ NP_2$  construction can be uniformly represented regardless of *no*’s syntactic property:  $\lambda y [NP_1(x) \wedge NP_2(y) \wedge R(x,y)]$ . Determining the relation *R* is crucial for (semantic) interpretation of  $NP_1 \ no \ NP_2$  constructions. In some cases, the interpretation of *R* depends on the semantic property of the *NPs*, which may have more than one possible meaning. In other cases, the interpretation of *R* may be contextually determined.

Adopting the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) as the framework for our analysis, we demonstrate, with a case study, how to derive the most plausible interpretations of *NPs* with *no* from contextual information.

**Keywords:** Japanese postposition *no*, context dependent interpretation, SDRT, Generative Lexicon

## 1. Introduction

Typologically, the Japanese language is an SOV language, and shows many of the grammatical features of this category. That is to say, the Japanese language is a head-last language, in which head constituents always appear at the end of any phrases. This language therefore uses postpositional, as opposed to prepositional, particles as follows:

(1) *Naomi ga Pari ni yuujin to it-ta*  
 Naomi NOM Paris GOAL friend with went  
 'Naomi went to Paris with her friends'

*No* is one of the most frequently used postpositions in Japanese. The meaning of this postposition corresponds roughly to the preposition "of" and the possessive marker "'s" in English.

The following phrases are examples of NP constructions containing *no*:

(2) *Naomi no haha*  
 Naomi mother  
 'Naomi's mother'

(3) *machi no hakai*  
 city destruction  
 'destruction of a/the city'

(4) *Toyota no kuruma*  
 Toyota car  
 'a car made by Toyota'

(5) *sencho no chichi*  
 captain father  
 'one's father, who is a captain,' 'a/the captain's father'

(6) *Sheekusupia no hon*  
 Shakespeare book  
 'Shakespeare's book,' 'a book on Shakespeare,' 'a book written by Shakespeare'

(7) *Pari no ie*  
 Paris house  
 'a house in Paris'

(8) *Naomi no chiimu*  
 Naomi team  
 'Naomi's team,' 'the team that Naomi expects to win,' etc.

(9) *San -nin no haha*  
 Three CL.person mother  
 'Three mothers,' 'the mother of the three (persons)'

Many scholars have already addressed how to interpret  $NP_1 no NP_2$  constructions. Shimazu, Naito, and Nomura's (1986) study represents a conventional approach: they classify nouns' semantic properties into several categories (e.g., animate, material, location, and others), classify semantic relations that stand between  $NP_1$  and  $NP_2$  (e.g., possession, whole-part relation, modification, etc.), and propose heuristics to construct the meaning of the whole NP from the constituents' meanings. Other researchers have analyzed the  $NP_1 no NP_2$  construction as an abbreviated form of related relative clauses. For example, Hirai and Kitahashi (1985) classify the semantic structures of noun phrases with relative clauses, and apply the same criteria to the analysis of  $NP_1 no NP_2$  constructions.

More importantly, however, every researcher concurs that an  $NP_1 no NP_2$  construction has a variety of meanings, and that its interpretation depends heavily on context information. Many scholars classify the meanings of the NP constructions with *no*, and attempt to formalize the method in order of deriving the overall meaning of an  $NP_1 no NP_2$  from the meanings of its constituents without recourse to context. However, there are few researchers that attempt to interpret these constructions using context information.

In this paper, we will conduct a syntactic and semantic analysis of *no*, and demonstrate, with a case study, how context information is crucial to the interpretation of NPs with *no*. For this study, we will adopt the Segmented Discourse Representation Theory (SDRT) of Asher and Lascarides (2003). Finally, we will discuss relevant topics such as how to interpret possessive expressions in English, and how to prepare necessary information to compute the meanings of NPs.

## 2. Syntactic Analysis of *no*

We employ Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and Japanese Phrase Structure Grammar (JPSG) (Sirai and Gunji, 1998; Gunji and Hasida, 1998) as the framework for our syntactic analysis. As we have shown elsewhere (Kikuchi, 2000), there are two kinds of *no* in  $NP_1 no NP_2$  constructions: one is a marker of complement (e.g., (2), (3)), and the other is the head of an adnominal (e.g., (4), (6), (7)).

In the former case, the  $NP_2$  noun phrase must have non-empty COMPS feature value. Typically, this type of  $NP_2$  is a predicative noun, which consists of a verb with the light verb *suru* (e.g., *hakai* 'destruction' + *suru* = *hakai-suru* 'destroy'), or a function/relation noun denoting an entity (e.g., *haha* 'mother,' *mae* 'front,' *ondo* 'temperature').<sup>1</sup>

In the latter case, *no*'s part of speech is adnominal, and it has non-empty ARG-ST (hence COMPS) feature value. We analyze typical possessive structure

such as *Naomi no neko* ‘Naomi’s cat’ as the case in point (cf. note 2). *Naomi no* is an adnominal phrase that modifies *neko* ‘cat.’

In the next section, we propose that the meaning of an  $NP_1 no NP_2$  construction can be uniformly represented regardless of *no*’s syntactic property. Even so, we claim that we should treat these two kinds of *no* separately. The main reason is that the complement marked by *no* appears as a gap in a long-distance dependency construction (e.g., (10), (11)), but the case of adnominal *no* (e.g., (12)) does not function in the same way:

(10) a. *ie no mae ga akichi -da*  
           house front NOM vacant COP  
           ‘(the) house’s front lot is vacant’

      b.  $[[\phi_i mae ga akichi -da] to Naomi ga shinji-teiru] ie_i$   
           front NOM vacant COP COMP Naomi NOM believe house  
           ‘(the) house whose front lot Naomi believes is vacant’

(11) a. *(kono) sakana no ryori ga taihen -da*  
           (this) fish cooking NOM hard COP  
           ‘(this) fish is hard to cook’

      b.  $[[\phi_i ryori ga taihen -da] to Naomi ga omot-teiru] sakana_i$   
           cooking NOM hard COP COMP Naomi NOM believe fish  
           ‘(the) fish which Naomi thinks is hard to cook’

(12) a. *Naomi no neko ga Taro ni kamitsui-ta*  
           Naomi cat NOM Taro GOAL bit  
           ‘Naomi’s cat bit Taro’

      b.  $??[\phi_i neko ga Taro ni kamitsui-ta] Naomi$   
           cat NOM Taro GOAL bit Naomi  
           ‘Naomi, whose cat bit Taro’

Since only the elements in ARG-ST feature value can be bound with its filler in a long-distance dependency, the above examples show that this kind of *no* is a real complement.<sup>2</sup>

### 3. Semantic Analysis of *No*

We need to keep in mind that the Japanese language has neither plural/singular markers nor indefinite/definite determiners. Consequently, in interpreting NP, there always exist ambiguities between definite and indefinite readings, and between plural and singular readings. Moreover, there are several ambiguities in the interpretation of  $NP_1 no NP_2$  constructions.

We hypothesize that, regardless of *no*’s syntactic behavior, the meaning of an  $NP_1 no NP_2$  construction can be represented as follows<sup>3</sup>:

### (13) Uniform Representation of an $NP_1$ *no* $NP_2$ Construction

$$\lambda y[NP_1(x) \wedge NP_2(y) \wedge R(x, y)]$$

Here,  $NP_1(x)$  and  $NP_2(y)$  represent that  $x$  is the object denoted by  $NP_1$  and  $y$  is the object denoted by  $NP_2$ .  $R(x, y)$  represents the relation between  $x$  and  $y$ .

In the case that *no* is a complement-marker, the meaning of *machi no hakai* ‘destruction of a/the city,’ for instance, is represented as follows:

$$(14) \lambda e [city(x) \wedge destroy(e, z, x)]$$

In this case, both  $NP_2$  and  $R$  are unified with *destroy*. That is, the ‘destroy’ event is described by the whole NP, and the destroyed object is represented by *machi* ‘city.’

Similarly, in the case that *no* is an adnominal, the meaning of *Naomi no neko* ‘Naomi’s cat,’ for instance, is indicated by the following:

$$(15) \lambda y [Naomi(x) \wedge cat(y) \wedge own(x, y)]$$

$NP_1$  and  $NP_2$  are replaced with *Naomi* and *cat*, respectively. Furthermore,  $R$  in (13) is replaced with a new relation *own*.

Kikuchi and Sirai (2002) classify the semantic patterns of  $NP_1$  *no*  $NP_2$  constructions according to how the relation  $R$  is derived:

- (A)  $NP_1$  largely determines the relation:  $NP_1$  is either a spatio-temporal location, which modifies  $NP_2$ , or a person/institution, to whom the referent of  $NP_2$  belongs (e.g., (7) and the possessive interpretation of (6)).
- (B)  $NP_2$  mainly determines the relation: If  $NP_2$  refers to an event, a relation, or a function, then the referent of  $NP_1$  functions as its argument. If  $NP_2$  refers to an object, then its qualia structure (Pustejovsky, 1995) determines the relation between  $NP_1$  and  $NP_2$  (e.g., (2), (3), and (4)).
- (C) Neither  $NP_1$  nor  $NP_2$  determines the relation. In some cases,  $R$  is contextually determined. In the other cases,  $R$  is equal, that is,  $NP_1$  and  $NP_2$  refer to the same object.

Although there are restrictions on the semantic properties of NPs that appear in (A) and (B) cases, typically there are several ambiguities in the interpretation of  $NP_1$  *no*  $NP_2$  constructions. Therefore, we have to determine from context information which interpretation is most plausible.

## 4. Framework of Context-Dependent Interpretation — SDRT

In the previous sections, we have demonstrated that  $NP_1$  *no*  $NP_2$  constructions are ambiguous both syntactically and semantically. Even if  $NP_2$  is a functional/relational noun,  $NP_1$  *no* is not necessarily its complement; it may be an adjunct, as in (5).

(5) *sencho no chichi*  
 captain father  
 'one's father, who is a captain,' 'a/the captain's father'

Furthermore, since the Japanese language lacks definite/indefinite determiners, we need context information in order to determine whether an NP denotes a specific individual or a non-specific one.

We adopt Asher and Lascarides' SDRT (Asher and Lascarides, 2003) as the framework for computing the pragmatically preferred interpretation of discourse. SDRT is an extension of (underspecified) Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) to include rhetorical relations such as *Narration* and *Parallel*. It distinguishes among several levels of discourse interpretation: lexical, syntactic (compositional), pragmatic, and cognitive levels.<sup>4</sup> Simply put, various ambiguities are represented with underspecifications in these levels. The discourse is represented as an SDRS, which is a recursive structure composed with labeled underspecified logical forms and rhetorical relations between the labels. It is updated with a new utterance (sentence) to include both old (the context's) and new (the current utterance's) information, to resolve some underspecified conditions, and to attain the maximally coherent interpretation. They propose the Maximized Discourse Coherence (MDC) principle in order to select the best updated discourse from possible updated discourses. As Asher and Lascarides (2003) have stated, the MDC principle captures the scalar coherence of a discourse interpretation, and essentially rests on the following rules:

- 1 All else being equal, the interpretation is more coherent if there are more rhetorical connections between two items in a discourse.
- 2 All else being equal, the quality of coherence of the interpretation is higher if there are more anaphoric expressions whose antecedents are resolved.
- 3 All else being equal, an interpretation which maximizes the quality of its rhetorical relations is more coherent than those that do not.

## 5. Case Study

In this section, we rely on the MDC principle to explain how to arrive at the most plausible interpretation(s) of  $NP_1$  *no*  $NP_2$  constructions. We will consider the following noun phrases: *sencho no chichi*, and *Taro no kuruma*.

### 5.1 Case where *R* depends on the existence of the NP's referent

As the first example, let's examine an interpretation of *sencho no chichi* where the relation *R* between *sencho* 'captain' and *chichi* 'father' depends on

whether the referent *sencho* or *chichi* is given in the context. The semantics of *sencho* is roughly represented as  $\lambda x[\text{captain}(x)]$ , and *chichi* is a relational noun that represents the father of some discourse entity. As we discussed in section 3, the semantics of *sencho no chichi* is represented as follows:

$$(16) \lambda y[\text{captain}(x) \wedge \text{father}(y, z) \wedge R(x, y)]$$

*Sencho no chichi* has (at least) two interpretations: ‘one’s father, who is a captain’ and ‘the captain’s father.’ These semantics are represented as (17) and (18), respectively:

$$(17) \lambda y[\text{captain}(x) \wedge \text{father}(y, z) \wedge = (x, y)]$$

$$(18) \lambda y[\text{captain}(x) \wedge \text{father}(y, x)]$$

$R(x, y)$  in (16) is replaced with  $= (x, y)$  in (17), and it is unified with  $\text{father}(y, x)$  in (18). In reality, it is difficult to get the latter reading *without appropriate context*. We can explain this fact in terms of the MDC principle. If the argument of *chichi* ‘father’ is identified, the referent (the father) is also identified. In other words, the referent of the whole phrase *sencho no chichi* is unidentified if the argument of *chichi* ‘father’ is unidentified. In (17), there is only one unidentified entity, the referent (the father), because its argument may be assumed to be the speaker according to the Japanese convention. In contrast, the speaker cannot be the candidate for the argument in (18). Therefore, there are two unidentified entities in (18), the referent and its argument. We can conclude from the MDC principle that this interpretation is less preferable to the first interpretation without appropriate context.

## 5.2 Case where $R$ is derived from context information

As the second example, let’s examine the two cases in which the relation  $R$  is derived from context information. Here, we use schema (13) from section 3 to represent the meaning of *Taro no kuruma* ‘Taro’s car’ as follows:

$$(19) \lambda y[Taro(x) \wedge \text{car}(y) \wedge R(x, y)]$$

Determining the relation  $R$  is crucial, and we will examine this phrase’s interpretation in two different discourses, one of which consists of (20.i) followed by (21), and the other consists of (20.ii) and (21).<sup>5</sup> Without any context, the most plausible interpretation of this phrase is a possessive reading (case A in section 3).

$$(20) \text{i. } Taro \text{ ga } jiko \text{ ni att-ta.}$$

Taro NOM accident DAT met

‘Taro had an accident’

$$\text{ii. } Taro \text{ ga } not-ta \text{ takushii } ga \text{ jiko } ni at-ta.$$

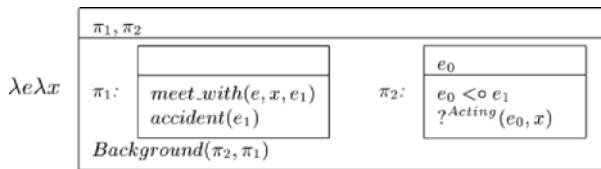
Taro NOM rode taxi NOM accident DAT met

‘The taxi that Taro rode had an accident’

(21) *Taro* no *kuruma* *ga* *koware-ta*.  
 Taro car NOM damaged  
 'Taro's car was damaged'

We assume that the lexical item *jiko* 'accident' denotes a complex event which is composed of two events,  $e_0$  and  $e_1$ , where  $e_1$  is the head event of this complex event and the non-head event  $e_0$  precedes and partially overlaps with  $e_1$ . We also assume that the semantic representation of *jiko ni atta* 'had an accident' can be illustrated by (22)<sup>6</sup>:

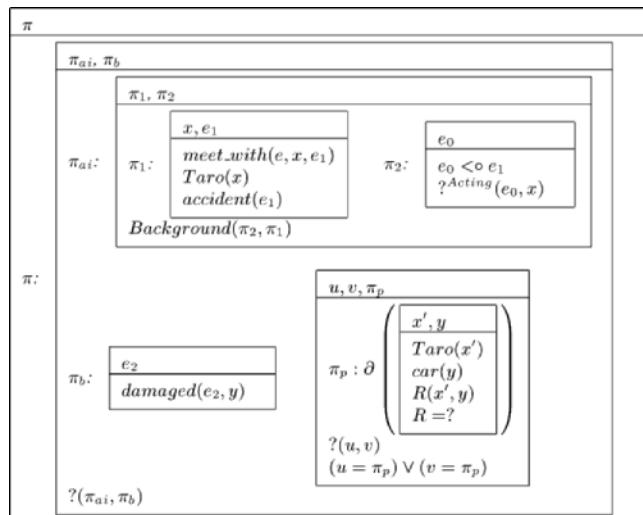
(22)



In the illustration, an agent  $x$  is doing an underspecified event  $e_0$ , which is shown by  $?Acting(e_0, x)$ , and he/she meets with an unfortunate event  $e_1$ .

We will first examine the interpretation of discourse (20.i)-(21) for which grammar produces (23) as the semantic representation (simplified):

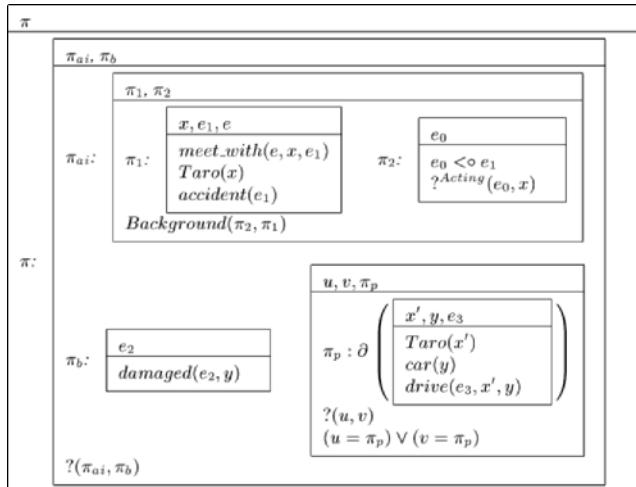
(23)



Here,  $\pi_{ai}$  and  $\pi_b$  are the labels of the logical forms for the utterances (20.i) and (21), respectively.  $R = ?$  means that  $R$  is underspecified, and  $\partial(\gamma)$  means that the content of logical form  $\gamma$  is presupposed. Moreover,  $?(\alpha, \beta)$  represents that there holds some rhetorical relation between  $\alpha$  and  $\beta$ . In the above illustration,  $?(u, v)$  and  $(u = \pi_p) \vee (v = \pi_p)$  represent that there is some rhetorical relation between  $\pi_p$  and some part of the context. Tense and aspectual information are ignored for simplicity.

Possible candidates for  $R$  in (23) are the possessive relation *own* (case A in section 3), and the event-relations derived from the qualia structures of the lexicon *kuruma* ‘car’ (case B). According to Generative Lexicon theory (Pustejovsky, 1995), we can assume that *drive* and *make* are designated in *kuruma*’s telic quale and agentive quale, respectively. Therefore, these event-relations are also the candidates for  $R$ . Replacing  $R$  in (23) with these candidates produces the possible updates of the discourse representations. We will show the result when  $R$  is unified with *drive* in (24):

(24)



In the case that  $R$  is unified with *drive* or *make*, the event  $e_3$  in  $\pi_p$  can be unified with the event  $e_0$  in  $\pi_2$ . This means that  $e_3$  precedes and partially overlaps with  $e_1$ . Furthermore, the *damaged* event is an unfortunate event which is

part of an accident. By the elaboration axiom schema (Asher and Lascarides, 2003), we can resolve  $?(π_{ai}, π_b)$  in (23) to be *Elaboration*( $π_{ai}, π_b$ ).

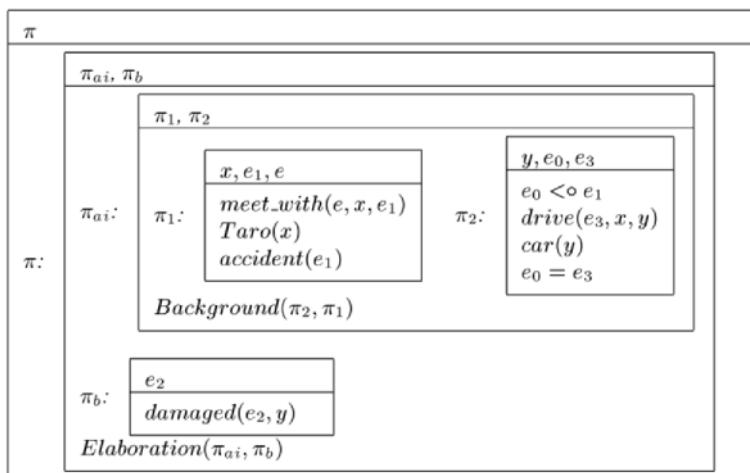
In contrast, considering the case of the possessive reading, the relation *own* cannot be unified with  $e_0$ , because it is not an event but a state. Therefore, by the MDC principle, *drive* and *make* are preferred to *own* as the candidates for  $R$ .

Moreover, if we apply the following axiom schema to this case, then we may conclude that *drive* reading is more plausible than *make* reading.<sup>7</sup>

$$(25) \quad (?(\alpha, \beta) \wedge [drive(e, x, y)](\alpha) \wedge [accident(e', x)](\beta)) > occassion(\alpha, \beta)$$

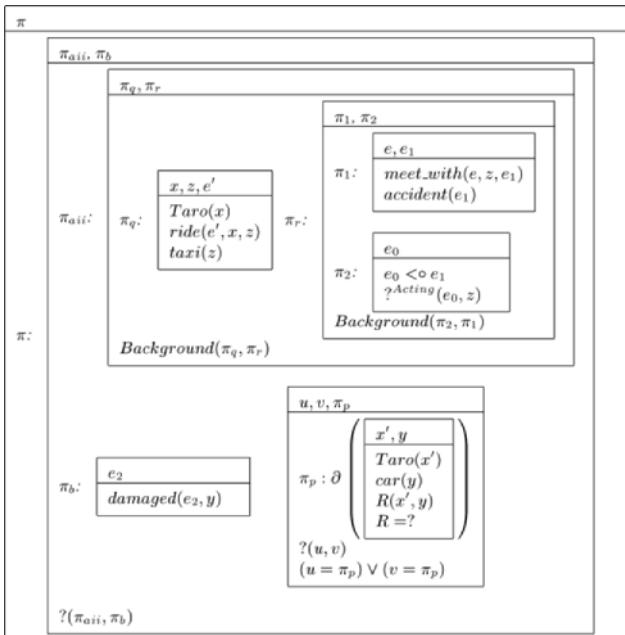
This means that if an agent( $x$ )'s driving and his accident are somehow connected, then the former occasioned the latter. Consequently, the *drive* reading is the most preferred interpretation for (20.i)-(21): 'Taro had an accident while he was driving. His car was damaged.' This matches with our intuition. Its semantic representation is shown in (26):

(26)



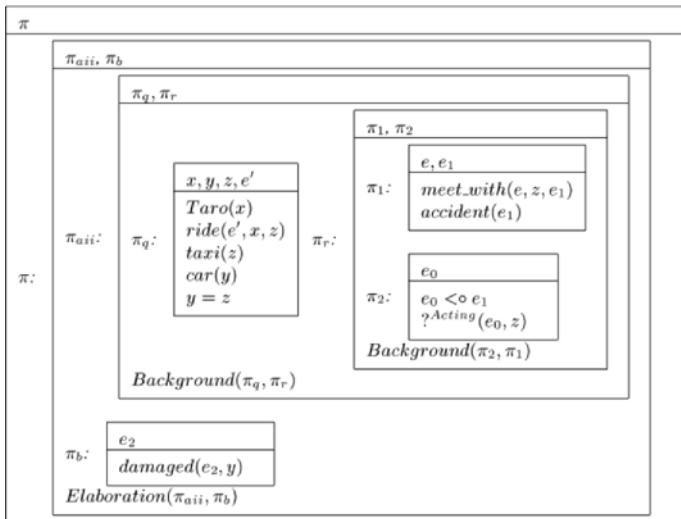
Now we will examine the interpretation of discourse (20.ii)-(21), which is shown in (27):

(27)



Here, the *noru* ‘ride’ event is explicitly specified as the relation between Taro and the taxi in (20.ii). Since a taxi is a type of car, the variables  $y$  in  $\pi_p$  and  $z$  in  $\pi_q$  can be unified, and *ride* and  $R$  share the same variables. Thus, we can infer that  $R$  is resolved to be *ride*. Moreover, we can induce *Elaboration* is the rhetorical relation between (20.ii) and (21) by the same method we used in our interpretation of (20.i)-(21). This produces the most preferred interpretation for this discourse, shown in (28): ‘Taro rode a taxi. His taxi was damaged.’ This interpretation also matches our intuition.

(28)



Considering the other candidates that resolve  $R$  to be *drive*, *make*, or *own*, the plausible interpretation should be the interpretation such that  $y$  in  $\pi_p$  is unified with  $z$  in  $\pi_q$ , and  $\pi_p$  is attached to  $\pi_{a(ii)}$  with *Background*. Among such interpretations, the *drive* reading is the most plausible one, but is less preferred to the *ride* reading shown in (28) by the MDC principle.<sup>8</sup>

## 6. Discussion

### 6.1 Evaluation

Table 1 shows the results of applying our method by hand to 2867 cases of  $NP_1$  *no*  $NP_2$  from the newspaper (Mainichi Newspapers, 1995). In this analysis, our method could successfully identify the most plausible interpretations in 2703 out of 2867 cases, or 94.3% of the time. In other words, our method was unable to select the most plausible interpretations in 164 of the

2867 cases.<sup>9</sup> This is not so bad comparing with 77% by Kurohashi and Sakai's (1999) method. In 937 cases among the correctly interpreted cases, the context information played the critical role in determining the relation  $R$ .

Table 1. Experimental result

correctly interpreted relations	94.3% (2703 / 2867)
incorrectly interpreted relations	5.7% (164 / 2867)

An example of incorrectly interpreted case of  $NP_1 no\ NP_2$  is the following:

(29) *kifu-kin*        *yaku 10-oku-en*        *no uchi*        *mottomo ookat-ta no*        *wa*  
       donated money    about 1 billion yen    among    the largest                                    TOP  
       *zenekon*        *nado*        *kensetsu-gyokai*        *no*        *yaku 3-oku-3-zen-man-en.*  
       general constructor etc.    construction industry        about 330 million yen.

'Among the donations, which amount to about 1 billion yen, the largest is the construction industry's 330 million yen.'

The problem is how to determine the relation between the construction industry and the 330 million yen. Without any contextual information, the most plausible interpretation would be a possessive reading. However, in this case, we can infer from the context that 'the construction industry, which includes general constructors and others, donated the largest amount, about 330 million yen.' We should know that '330 million yen' is 'a part of donated money' from the meaning of the sentence itself. Thus, we can conclude that '330 million yen' from this information that '330 million yen' is 'donated money,' and that there should exist a donor and a recipient. Consequently, we may reason that the 'construction industry' is the donor. Currently, we are not certain that this kind of inference is allowed in the glue logic of SDRT. We will have to investigate similar phenomena.

## 6.2 Comparison with other works

Partee and Borschev (2001) examine possessive constructions in English. They claim that there are two kinds of possessive constructions: argument-type and modifier-type. They also claim that the interpretation of the possessive construction is determined by either the head noun's inherent relation or the contextually dependent relation  $freeR$ . Vikner and Jensen (2002) argue that  $freeR$  can be divided into two types: one is truly context dependent, and the other is derived from the qualia structures (Pustejovsky, 1995) of NPs. They also present a method to predict the possible lexical interpretations of prenominal genitive constructions. Their analysis of English possessive constructions looks very similar to our analysis of the Japanese postposition *no*, but they do not discuss how to determine the most plausible interpretation, in particular,

how *free R* is resolved. We think that, although the range of possible relations denoted by English *free R* may be different from that of *no* in the Japanese language, our approach can be used to determine the most plausible interpretation of English possessive constructions.

### 6.3 Realization of our method

There are several problems in implementing our approach. First of all, we need an ontology that includes the type hierarchy of lexical items containing lexical knowledge, especially information of qualia structures. We may have to extend the current electronic dictionaries (Fellbaum, 1998; Japan Electronic Dictionary Research Institute, 1993; Ikehara *et al.*, 1997) by exploiting the ideas of Buitelaar (1998) and Lapata and Lascarides (2003) to build the desirable ontology.

Another problem is determining whether axiom schema (25) from section 5.2 is necessary. We applied this axiom to discourse (20.i) and (21) and conclude that the *drive* relation is more plausible as the candidate for *R* than the *make* relation. But as this axiom seems a little ad hoc, we may need an alternative way to make the *drive* interpretation more plausible than the *make* interpretation. Pustejovsky (1998) defines the agentive qualia as follows:

$$(30) \quad \text{Agentive}(\lambda x[\alpha(x)]) = \lambda e(\psi(e) \leftrightarrow \forall x \forall e[\alpha(x, e) \rightarrow \exists e' \exists y[\psi(e') \wedge e' \prec e \wedge \text{make}(e', y, x)]])$$

According to this definition, the *make* event, which is derived from the agentive quale of *kuruma* ‘car,’ must precede the *accident* event. Thus, the *make* event cannot be unified with the underspecified event  $e_0$  in (22). Consequently, we can explain the reason why the *drive* reading is preferred to the *make* reading without using axiom schemata such as (25).

Third problem comes from the fact that we use default reasoning to resolve rhetorical relations among clauses and sentences. In order to utilize default reasoning, we need a consistency check, which had been thought unworkable. That is one of the reasons why we adopt SDRT as our framework, because it separates logics of information into several domains, and the domain where we need a consistency check is restricted to propositional logic. Thus we think that our method is theoretically reasonable and realizable even if we use default reasoning.

## 7. Conclusion

We have proposed a syntactic and semantic analysis of the Japanese postposition *no*, and analyzed the possible interpretations of  $\text{NP}_1 \text{ no } \text{NP}_2$  constructions. We have shown how to determine the most preferable interpretation using the SDRT framework with the MDC principle. We think that we can ap-

ply this approach to the analysis of compound nouns, adnominal constructions such as relative clauses, and other postpositions.

## Notes

1. These nouns may take as their complements PP as well as NP marked by *no*. Other nouns that take as their complements NP and PP marked by *no* can be shown in this way:

(31) (*Naomi* no) *Pari kara* no *tegami*  
 Naomi      Paris from      letter  
 '(Naomi's) letter from Paris'

Compare this with the following, which lacks *no*:

(32) \**Pari kara tegami*  
 Paris from letter  
 'a letter from Paris'

2. There are two types of *no*, when they appear in possessive constructions: adnominal and complement-marker. For example, the *no* in *zou no hana* 'elephant's trunk' is a complement marker, because this allows a long-distance dependency as follows:

(33) [[ $\phi_i$  *hana*      *ga*      *nagai koto*] *o*      *Naomi* *ga*      *shit-teiru*] *doobutsu*  
 trunk/nose    NOM    long    COMP    ACC    Naomi    NOM    know    animal  
 'an animal whose trunk Naomi knows to be long'

Actually, we think of *hana* 'trunk/nose' as a functional noun, because this is an untransferable, dependent part of the possessor.

3. In this paper, we do not take such cases as *nise no kane* 'fake gold' into account. Since 'fake gold' is not real gold, (13) does not hold (i.e.,  $\exists x$  fake-gold(*x*) does not entail  $\exists x$  gold(*x*)).

4. For the sake of simplicity, we ignore the cognitive level in this paper.

5. There are some native speakers who claim that *Kuruma ga kowareta* is better than (21). Even so, they will admit that this *kuruma* 'car' means *Taro no kuruma* 'Taro's car'.

6. This is analogous to Asher and Lascarides' idea that some nouns have rhetorical relations specified as part of their lexical semantics (Asher and Lascarides, 2003).

7. Not only lexical knowledge but also some world knowledge may be needed to infer this axiom schema.

8. It is not clear whether we can obtain the most plausible interpretations in this case by the original MDC principle. Here, we use the modified MDC principle with the following additional rule: the quality of coherence of the interpretation is higher if the label for the presupposed information is anaphorically bound (i.e., identical) with an available label than it is related by rhetorical relations.

9. In sixteen cases out of the incorrectly interpreted cases, which are names of books or artifacts, even human beings had difficulty in determining the relation *R* without domain knowledge.

## References

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Buitelaar, P. (1998). *CoreLex: Systematic Polysemy and Underspecification*. PhD thesis, Brandeis University.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gunji, T. and Hasida, K. (eds.) (1998). *Topics in Constraint-Based Grammar of Japanese*. Kluwer.

Hirai, M. and Kitahashi, T. (1985). Nihongobun ni okeru *no* to rentai-shushoku no bunrui to kaiseki (A semantic classification of noun modification in Japanese)

anese sentences and their analysis). *Johoshori Gakkai Shizengengoshori Kenkyukai (IPSJ SIG-NL)*, 58(1):1–8. (in Japanese).

Ikehara, S., Miyazaki, M., Shirai, S., et al. (eds.) (1997). *Nihongo Goi Taikei (A Japanese Lexicon)*. Iwanami Shoten. (in Japanese).

Japan Electronic Dictionary Research Institute (ed.) (1993). *Gainen Jisho (The Concept Dictionary)*. Japan Electronic Dictionary Research Institute.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer.

Kikuchi, R. (2000). Goi imi joho ni motozuku nihongo meishiku no imikaiseki — “A no B” o rei ni (Lexical information based semantic analysis of Japanese noun phrases: A case of “A no B”). Master’s thesis, Graduate School of Computer and Cognitive Sciences, Chukyo University. (in Japanese).

Kikuchi, R. and Sirai, H. (2002). Nihongo meishiku no imikaiseki no kento (Study of Japanese noun phrase semantic analysis). In *Proceedings of 19th Annual meeting of the Japanese Cognitive Science Society*, pp. 134–135. (in Japanese).

Kurohashi, S. and Sakai, Y. (1999). Semantic analysis of Japanese noun phrases: A new approach to dictionary-based understanding. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, pp. 481–488.

Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):263–317.

Mainichi Newspapers (1995). CD-Mainichi Shimbun ’95. Nichigai Associates.

Partee, B. H. and Borschev, V. (2001). Some puzzles of predicate possessives. In Kenesei, I. and Harnish, R. M. (eds.), *Perspectives on Semantics, Pragmatics, and Discourse*, pp. 91–117. John Benjamins Publishing.

Pollard, C. and Sag, I. (eds.) (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Pustejovsky, J. (1998). The semantics of lexical underspecification. *Folia Linguistica*, XXXII.

Shimazu, A., Naito, S., and Nomura, H. (1986). Joshi no ga musubu meishi no imikankei no kaiseki (Analysis of semantic relations between nouns connected by a Japanese particle *no*). *Keiryo kokugogaku (Mathematical Linguistics)*, 15(7):247–266. (in Japanese).

Sirai, H. and Gunji, T. (1998). Relative clauses and adnominal clauses. In Gunji, T. and Hasida, K. (eds.), *Topics in Constraint-based Grammar of Japanese*, pp. 17–38. Kluwer.

Vikner, C. and Jensen, A. (2002). A semantic analysis of the English genitive. Interaction of lexical and formal semantics. *Studia Linguistica*, 56(2): 191–226.

## Chapter 17

# WHAT DO THE NOTIONS OF INSTRUMENTALITY AND OF MANNER HAVE IN COMMON?

*A Channel Theoretic model for causality as dependence*

Alda Mari

CNRS - ENST

46, rue Barrault - 75013 Paris - F

mari@enst.fr

**Abstract** In this paper we present an analysis and a model for the notions of instrumentality and manner through the study of the preposition *avec* (*with*).

These two notions are very often assimilated. We try to find out the semantic foundations of this intuition, considering first the meanings-in-context and then the underspecified representation common to them.

For the construction  $NP_1 VP NP_2$  *avec*  $NP_3$ , we propose an analysis in terms of sub-events involving the denotations  $NP_1$  and  $NP_3$  and individuate two features shared by these meanings:

1. the causal relation linking the individual entities among themselves and with respect to the main action;
2. the situation dependence of these relations.

We propose a model based on the abstract notions of type, constraints and channel, which allows us to capture the abstract notion of causation as non-accidental association or property dependence ((Lewis, 1973)).

**Keywords:** Instrumentality, manner, *avec* (*with*), causality, situation, types, counterfactuals, dependence, accidentality, underspecification.

### 1. Aim and methodology

It has long been recognized that instrument and manner are two primitive, conceptually related notions that are generally introduced by a prepositional

phrase ((Spang-Hanssen, 1963); (Wierzbicka, 1996); (Cadiot, 1997)):

- (1)  $NP_1 VP \ NP_2 \ Prep \ NP_3$
- (1')  $NP_1 VP \ NP_2 \ avec \ NP_3$
- (2) Jean coupe le pain avec un couteau  
*John cuts the bread with a knife*
- (3) Jean joue aux échecs avec plaisir  
*John plays chess with pleasure*

*Prep*  $NP_3$  is supposed to specify what way in particular the action denoted by the *VP* is carried out, introducing a qualification in the scenario enhanced by the verb. In (2), *avec un couteau* (*with a knife*) makes explicit the way of cutting the bread; in (3), *avec plaisir* (*with pleasure*) qualifies the way of playing.

In this paper we argue that this intuition is well founded, and we try to discern its foundations. To this end we analyze and model the notions of instrumentality and manner through the study of the preposition *avec* (*with*) in French. More than *par* (*by*), which instantiates a general meaning of cause or “way through”, *avec* - as its literal translations in other Indo-European languages (*con* it., *mit* ger., *with* engl.) - has, among many others (see ((Mari, 2003))), the meanings of instrument and of manner for which (2) and (3) are two typical examples.

Our study is based on the hypothesis that if primitive notions such as instrument and manner exist, they can only be studied through their possible lexicalizations. Moreover, because the meanings of an item can only be observed in context, the explanation of the similarity between these two notions follows from a bottom-up analysis, from the meanings to the abstract representation.

This paper is then structured as follows: we first describe the contextual behavior of *avec*-instrument and *avec*-manner in section 2. We take into account some features such as the distribution of determiners, the question of the  $NP_3$  types and the constraints on predicate interpolation. Then, we present a model based on the notion of type and property constraint inspired by ((Barwise and Seligman, 1997)) (section 3). In section 4 we develop a representation for the underspecified scheme that the notions of instrument and manner share, and we finally come back to a formal treatment of *avec* contextual meanings. We conclude (section 5) with an evaluation of our results and some theoretical speculations. But, to begin, let us present a brief note on *avec* and our methodology.

## 1.1 A note on *avec* and the scope of this study

The strategy that we adopt consists in observing the notions of instrument and manner in their maximal proximity, that is to say, in the configuration where they are instantiated by the same lexical item.

This strategy raises two questions:

1. are these two notions only lexically driven or,
2. can we extend the conclusions to other cases and show that their similarities are lexically independent?

It is well known that *avec* is a highly polysemous preposition. Among others, it has the meaning of comitativity “Jean is walking with Mary” (Mari, 2002) and of influence “John is watching the TV with his brother singing next to him”. Instrumentality and manner are two specific senses of *avec*. A complete study (Mari, 2003) of the meanings of *avec* shows that they belong to the same sub-family of *spatio-temporal location*. In this paper we concentrate on this subset and we show that the notions of instrumentality and manner are related: there is an underspecified representation that these meanings share. Specific parameters and constraints instantiate it in context ((Pinkal, 1985); (Poesio, 1996)).

To show that instrumentality and manner are two related, lexically independent notions, we would have to tackle some other items that present the meaning of manner and, separately, some other items or construals that present the meaning of instrument, and then show that these two meanings are related, an endeavor which is beyond the scope of this paper.

Nevertheless, the nature of the representation that we adopt, based on the notions of constraints and situation type ((Barwise and Perry, 1983) ; (Devlin, 1991) ; (Barwise and Seligman, 1997)) leads us to formulate the hypothesis that the similarities we have found extend beyond the lexical meaning of the preposition *avec*.

## 2. Analysis of *avec*-instrument and *avec*-manner

In this section we consider the meanings of instrument and manner, and we describe the lexical constraints that must be fulfilled for the use of *avec*. In the following discussion,  $NP_1$ ,  $NP_2$ ,  $NP_3$  and  $VP$  refer to construction (1');  $X$ ,  $Y$ ,  $Z$  are, respectively, the denotations of  $NP_1$ ,  $NP_2$  and  $NP_3$ <sup>1</sup>.

### 2.1 *Avec*-instrument

Let us begin by considering the features that characterize the meaning of instrumentality. Recall that the typical example is:

(2) Jean coupe le pain avec un couteau

*John cuts the bread with a knife*

**2.1.1 Causality.** As shown in (Mari and Saint-Dizier, 2003), the notion of instrument is not self-standing. It cannot be found at the level of the  $NP_3$  type, it is not required as such by a particular class of verbs, nor does it have a unique lexical and conceptual representation. Instead, it minimally involves three sub-events and some relations among them: ( $e_1$ )  $X$  carrying out the action denoted by  $VP$ , ( $e_2$ )  $X$  controlling or undergoing  $Z$ , and ( $e_3$ )  $Z$  causing the action described by the  $VP$ .

**(4) Sub-events: analysis of *avec*-instrument (2)**

- $e_1$  : Jean cuts the bread ( $X$  / *Action*);
- $e_2$  : John uses the knife ( $X$  “controls”  $Z$ );
- $e_3$  : The knife cuts the bread ( $Z$  / *Action*).

Talmy (Talmy, 1976) calls the reconstructed event ( $e_3$ ) denoted by  $NP_3$   $VP$  the “semantic cause”. It is *because* the knife has a certain property that enables it to cut the bread (e.g. its sharpness), that John can cut the bread. Of course, there must exist a relation between John and the knife. This is provided by sub-event,  $e_2$ . Different prepositions denoting instrumentality signal different degrees of involvement of  $X$  and  $Z$  in the action denoted by the  $VP$ , as well as different control degrees of  $X$  on  $Z$  (cf. Mari and Saint-Dizier, *ibid.*). On its side, *avec*-instrument shows a particular configuration of control relations.

**Relation  $X$  /  $Z$ .**  $X$  must control  $Z$ . This control can be physical (5), psychological (6), or even intellectual (7):

(5) Jean a cassé la fenêtre avec une balle

*John broke the window with a ball*

(6) Jean a évité la réunion avec une excuse peu vraisemblable

*John avoided the meeting with an unbelievable excuse*

(7) Le gouvernement a évité la manifestation avec une longe déclaration bien sentie

*The government avoided the demonstration with a long heartfelt declaration*

Note that the following sentences are impossible:

(8) \*Jean a mis tout le monde mal à l'aise avec une gaffe

*John embarrassed everyone with a gaffe*

(9) \*Jean s'est fait mal avec une chute

*John hurt himself with a fall*

*Gaffe* and *chute* are non-controllable entities. Note also that the control need not be physical (5) vs. (6).

**Relation X / Z / Action.** It does not follow that the subject performs the action denoted by the *VP* voluntarily. In (5) the fact of breaking the window can be the unwanted result of a bad control of the ball. In other cases, the subject is volitional. According to Talmy's distinction (Talmy, *ibid.*) between *actor* (non volitional agent) and *agent* (a volitional actor), we can state that *X* can be either actor or agent. From these two observations, we can conclude that in (1'):

- *X* has to *control* some of the properties of *Z* thus causing the action denoted by the *VP* even though involuntarily<sup>2</sup>
- there is a causal relation between the properties of *Z* and the main action.

### 2.1.2 Situation-dependence of the properties of the instrument.

Another characteristic of *avec*-instrument is that the properties of *Z* are situation-dependent.

(10) Jean épate Marie avec sa voiture

*John is impressing Mary with his car*

If uttered in the 19<sup>th</sup> century, the existence of the car itself would have been sufficient to impress anyone. If uttered nowadays, the interpretation of the sentence would depend on the reconstruction of some relevant properties of the car that impress Mary. Moreover, the fact that the properties relevant for the interpretation are situation-dependent explains two other observations.

**Lexical and contextual constraints for the reconstruction of the interpolated predicate.** The relation between *NP*<sub>3</sub> and the *VP* takes the form of an interpolated predicate. In (4) the knife "cuts" the bread, in (10) the declaration "convinces" the public opinion and so on. With respect to the meaning of instrument, one could expect to pick the material to interpolate from the telic role of the *NP*<sub>3</sub> (Pustejovsky, 1995). Nevertheless, even if this option is the one by default, in most of the cases, the predicate to interpolate is provided by the lexical information and the context (Godard and Jayez, 1993). In (11), for instance,

(11) Il a écrasé les moustiques avec un livre / *He swatted the mosquitoes with a book*

the predicate to reconstruct is not “to read,” as expected by default. Other properties of the book are called into play, namely the physical property of being heavy enough to swat the mosquitoes.

### Determiners and conditions on situation-dependence of properties.

Some kinds of determiners are not compatible with *avec*-instrument. Consider the following examples.

(12) \*Le gouvernement a calmé les manifestants avec la déclaration  
*??The government has calmed down the demonstrators with the declaration*<sup>3</sup>

(13) Le gouvernement a évité les manifestations avec la déclaration et calmé les parlementaires avec la promulgation de la nouvelle loi  
*The government avoided the demonstrations with the declaration and calmed down the Chamber with the promulgation of the new law*

(14) Le gouvernement a calmé les manifestants avec la déclaration que le Président a prononcée hier  
*The government calmed down the demonstrators with the declaration that the President gave yesterday*

(15) \*Le gouvernement a calmé la manifestation avec la longue déclaration  
*\*The government calmed the demonstrations with the long declaration*

(16) Hier, le gouvernement a préparé une longue déclaration. \*Aujourd’hui il a calmé les manifestations avec la déclaration  
*Yesterday the government prepared a long declaration. \*Today it calmed the demonstrations with the declaration*

(17) Le gouvernement a calmé les gens avec sa / cette déclaration  
*The government has calmed down the people with its / this declaration*

(18) Le gouvernement a calmé les manifestations avec une déclaration  
*The government has calmed the demonstrations with a declaration*

First of all, the definite is not acceptable when the properties of the instrument are not salient in the context (12). Following ((Corblin, 1995)), for the salient properties to emerge, a contrast has to be established (13). Eventually, the definite *NP* has to be spatio-temporally anchored (14). The overt specification of the relevant properties by an adjective is not sufficient (15), nor to know, on the basis of the context, the properties of the denotation that make it worth mentioning (16). Because the possessive and the demonstrative are context-anchored determiners, they are always possible (17). The indefinite, introducing a new entity into the context, is more easily accepted, especially when this is anaphorically bound by another salient entity as in (18).

We conclude that the properties of the instrument that allow *X* to achieve the action described by the *VP* are situation dependent and have to be relevant in the situation.

## 2.2 *Avec-manner*

There are many resemblances between *avec-instrument* and *avec-manner*. The most straightforward ‘ is that the subject, in construction (1') for *avec-manner*, also has to minimally be an actor (Talmy, 1976).

(3) Jean joue aux échecs avec plaisir

*John plays chess with pleasure*

(19) \*Jean plaît avec enthousiasm

*\*John is liked with enthusiasm*

The  $NP_3$  has to qualify the action carried out by *X* in the way *avec-instrument* does. This explains why it has very often been said that the meanings of manner and instrument can only be distinguished on the basis of the type of the head noun of  $NP_3$ . *Avec-instrument* would require a concrete object, where *avec-manner* would select an abstract one.

We argue that the *resemblances* (Wittgenstein, 1953) cannot be reduced to this simple rule. First of all, as we show below, the nouns that can appear in construction (1') for *avec-manner* are not clearly defined for their type. On the other hand, the similarities concern the structuring relations among the *NPs* and the *VP* rather than the content of the *NPs* alone.

Our analysis of *avec-manner* shows that its two major characteristics are, as for *avec-instrument*, the situation-dependence of the properties of *Z*, and the causal relation that links these properties to the action described by the *VP*.

But before we come to these two points, we discuss two of the major theories of *avec-manner* in French and show that they are not explanatory adequate.

**2.2.1 *Avec-manner* and *NP* types.** As for the instrumental, one would try to look for the notion of manner into the  $NP_3$  type. Given the apparent coherence of the semantics of the head nouns, this hypothesis seems reasonable. They are generally abstract items, belonging to the classes of *intensive* (roughly mass terms) and *extensive* (roughly countable terms) nouns (Flaux and van de Velde, 1993). Nevertheless, it is very difficult to identify a proper set that can enter construction (1') on the basis of distributional properties. This enterprise has been undertaken by (Molinier, 1984). The only result is that all the nouns that can enter construction (1') can be the object of *éprouver* (*to feel*) or *manifester* (*to show*). Either syntactically or semantically this result does not clarify what semantic properties make these nouns acceptable as complements of *avec-manner*. We have in fact to note that

a certain number of abstract nouns are not compatible with this preposition:

(20) Avec \**beauté* (*beauty*), \**solitude* (*loneliness*), \**célébrité* (*celebrity*), \**silence* (*silence*)

Both the semantic and the syntax-semantic accounts seem unable to provide a plausible explanation. Both of them try to identify specific and permanent properties in the denotations of the nouns without considering the constraints imposed by the preposition. Let us consider them in turn.

**Semantic account.** Anscombe (1990) has provided a classification of abstract nouns on the basis of the nature of the property that they express. He proposes a distinction between *intrinsic* (*gentillesse* (*kindness*), *blondeur* (*blondness*)) and *extrinsic* (*intérêt* (*interest*), *silence* (*silence*)) properties. The intrinsic properties (without necessarily being permanent) characterize an individual as such; the extrinsic ones describe transitory states. It is easy to observe that *avec* is compatible and incompatible with items belonging to both of these classes:

(21) a. Intrinsic : *avec gentillesse* / \**avec blondeur* / b. Extrinsic : *avec intérêt* / \**avec silence*

The second distinction established by Anscombe (*ibid.*) is between *endogenous* (*courage* (*courage*)) vs. *exogenous* (*méfiance* (*skepticism*)) properties. The first are supposed to have an internal psychological source; the second are considered to be enhanced by an external element. The status of this characterization is quite vague and difficult to state on a semantic basis. Moreover *avec courage* and *avec méfiance* are both possible.

We conclude that none of these conceptual distinctions help us to identify the nouns that can combine with *avec*-manner.

**Syntax-semantic account.** Some of the nouns that cannot enter construction (1') in the *NP*<sub>3</sub> position belong to the class of state-nouns (Flaux and Van de Velde, *ibid.*). This class is identified by the context: *être en* (literally *to be in*).

(22) a. *Etre en colère* (*anger*), *désordre* (*untidiness*) / b. *avec ??colère*, *avec \*désordre*

This is a promising hint for the analysis of *avec*-manner. However, we have to acknowledge what follows:

- first of all, *avec* is also incompatible with some abstract nouns that are non-state nouns:

(23) a. \*Etre en volonté (*to be in will*) / b. avec \*volonté

- secondly, the semantics of *être en* is uncertain. Leeman (Leeman, 1995) argues that only episodic resultative nouns can follow the preposition *en* (*in*), without making clear what exactly a resultative noun is; moreover, Leeman's conclusion contradicts the account of (Flaux and Van de Velde, *ibid.*) and the notion of permanent state.

We conclude that this account only represents a tiny hint toward the individuation of the properties that a generally abstract noun has to fulfill in order to enter construction (1').

**2.2.2 Situation dependence.** We suggest that a noun is a possible candidate for construction (1') in  $NP_3$  position if the properties it expresses are situation-dependent. In this respect, the distinction between individual-level predicates and stage-level predicates seems to better capture the data (Carlson, 1977) : only stage-level predicates are acceptable. It is well known that intrinsic properties belong to the individual-level. These are detected by the classical test using perception verbs:

(24) a. J'ai vu Jean \*beau - *I have seen John \*handsome* / b. avec \*beauté - *with \*beauty*

Nevertheless, this distinction is not sufficient to determine the candidates for  $NP_3$ . Some of the nouns, even if they denote stage-level predicates, cannot follow *avec*-manner:

(25) \*Jean regardait la télévision avec solitude  
*\*John was watching the TV with loneliness*

**2.2.3 Causal dependence of the property in the scene denoted by the VP.** The denotation of  $NP_3$  also has to be causally related to the *VP*. Let us briefly discuss the notion of causal relation.

(26) \*Jean regarde la télévision avec dépression  
*\*John watches TV with depression*

Even if *dépression* can denote an episodic property of John, this property is thought of as independent of the action of watching the TV. This explains why (26) cannot be accepted.

In (3), instead, “pleasure” is linked to the act of playing chess.

This observations points to the fact that the association between the eventuality denoted by the predicate and the property denoted by  $NP_3$  need *not be accidental*. These properties have to be related to each other, or dependent. Note that it is not the case that the “pleasure” causes the fact that John plays chess. Again, the notion we are pointing at, is more that of “property dependence.” The fact of feeling “pleasure” is dependent on the action of playing chess. We call this dependence “causal dependence” (see ((Lewis, 1973)) and the notion of counterfactuality).

Even though it may seem that the notion of cause has to be captured more abstractly for *avec*-manner than for *avec*-instrument, as we develop it in detail at section 3., the very same notion of regularity and non-accidental link (Lewis, 1973) is involved in both of the cases.

## 2.2.4 *Avec-manner* and its determiners.

## The existence of a

causal dependence between the property expressed by  $NP_3$  and the eventuality/situation described by the  $VP$  is confirmed by the analysis of the determiners.

- *Determiner “zero”* ( $\emptyset$ ). Following (Anscombe, 1990) for the analysis of construction (27):

(27)  $VP$   $N_i$  or  $P$   $N_i$  or  $N$   $N_i$

we assume that *determiner* in French is meaningful and indicates that there is a temporal and causal coincidence between the action denoted by the *VP / P / N* and the property denoted by  $N_i$ . In this configuration,  $N_i$  denotes a property that *structurally* describes the predicate. For instance in *confiture pur sucre (marmalade pure sugar)*, *pure sugar* qualifies a kind of marmalade resulting from a particular treatment. In this case, a causal link exists between the substance and its property: given a certain process of production, it is not accidental that the resulting substance is “marmalade pure sugar”<sup>4</sup>. What Anscombe (*ibid.*) seems to mean by “causal link”, then, is an ontological dependence between the property being marmalade and being pure sugar.

This analysis of determiner “zero” confirms our hypothesis that in construction (1’), *Z* preceded by the determiner has to be causally involved in the scenario enhanced by the *VP*. To put it otherwise, there has to be a non-accidental association between the property describing the eventuality denoted by the *VP* and *Z*. The *non-accidental association* is nothing but an *ontological dependence*, or a *causal relation* between two properties.

A property that is dependent on the substance is called *trope* (Simons, 1994). That the  $NP_3$  denotes a trope seems to be a general requirement.

- *Tropes*. A trope is syntactically obtained by the introduction of a modifier at the syntactic level.

(28) \*Jean a accueilli Marie avec la joie

\*John has welcomed Mary with the joy

(28') Jean a accueilli Marie avec la joie au coeur / la joie d'un vrai ami

(literally: John has welcomed Mary with (?the) joy in his heart / the joy of a true friend)

The fact that the presence of modifiers is mandatory whenever the property has to be made situation dependent leads us to conclude that *avec*-manner requires that this move be made for the sentence to be felicitous.

We can then claim that:

- the simultaneity between the realization of the property denoted by the  $NP_3$  and the action denoted by the *VP* is a necessary but not a sufficient condition;
- *avec*-manner requires a causal link between *Z* and the action described by the *VP*.

These observations lead us to the conclusion that *avec*-manner enters a structure of three sub-events:

**(29) Sub-events: analysis of *avec*-manner**

- $e_1$ : *X* performs the action denoted by the *VP*;
- $e_2$ : *Z* is causally linked to (i.e. non-accidentally related to, or dependent on) the action denoted by the *VP*;
- $e_3$ : *X* is the source of (generally “feels”) *Z*.

This tri-partition can be easily and straightforwardly compared to the one given for *avec*-instrument above (4).

## 2.3 Conclusion of the analysis

We can conclude our analysis of *avec*-instrument and manner by stating that they can both be analyzed in terms of sub-events. Abstracting from the representations in (4) and (29) the common features of these sub-events are the following:

### (30) Sub-events: analysis of *avec*-instrument/ manner and situation dependence

- $e_1$ :  $X$  performs the action denoted by the *VP*;
- $e_2$ :  $Z$  is causally linked (i.e. non-accidentally related to, or dependent on, ((Lewis, 1973)) to the action denoted by the *VP*;
- $e_3$ :  $X$  and  $Z$  are in a certain relation (control, or psychological source relation) with respect to the action denoted by the *VP* (*situation dependence of the relation*).

The content of  $e_3$  follows from  $e_1$  and  $e_2$ . It can be paraphrased in the following way: Construction (1') denotes a scene in which  $X$  and  $Z$  are two entities whose properties are related. What clearly differentiates *avec*-instrument from *avec*-manner is the nature of the relation between  $X$  and  $Z$  ( $e_3$ ): in the first case this takes the form of a control, in the second case  $X$  is the psychological source of  $Z$ . This relation exists by virtue of the existence of a unique involving situation. On the other hand, this involving situation exists because  $X$  and  $Z$  are related: the bread can be cut with the knife because John uses the knife which has a certain property that enables it to cut (e.g. its sharpness). Chess is “played with pleasure” because “John feels pleasure while playing chess.”

Let us emphasize again, as we have mentioned above, it is not the case that the pleasure is the “cause” of the fact that John plays checks. The notion of “causal relation” stays at a more abstract level and amounts to that of “ontological dependence” of properties, or “non-accidental association.” It is the case that feeling pleasure is dependent on the eventuality of playing chess and reciprocally, the eventuality of playing chess intrinsically involves that of feeling pleasure. This coordination of properties ((Lewis, 1973)) is the abstract causal link that constitutes the common core of the meanings of instrumentality and manner, as they are instantiated by *avec*.

### 3. The model: properties and constraints

The model we have developed to explain the behavior of *avec*-instrument and *avec*-manner is inspired by Channel Theory of ((Barwise and Seligman, 1997)). This is a theory of distributed systems: wholes, whose parts have a co-ordinated behavior. Consequently, the theory is not used as a mere formalism, but as a model whose expressive power is entirely exploited.

Let us very briefly introduce the main definitions and emphasize their relation with the issues developed so far. No other acquaintance with the theory is required in order to read this section.

- **Objects.** Ordinary objects or entities (e.g. tables, individuals, animals etc.), properties (e.g. blondness, patience, etc.) or eventualities<sup>5</sup> are *Objects*. They can all be described by, at least, their spatio-temporal location.

- **Types.** Types are descriptions of *Objects*. Technically, it has to be possible to assign at least one type or description to each object in a given situation. From a semantic point of view, we can consider types as *trope*s denoting spatio-temporally anchored properties of entities.

(31) **Classification.** A classification is a triple  $(Objects, Types, \models)$ , where *Objects* is a set of objects, *Types* a set of categories or types, and  $\models$  a relation between *Objects* and *Types*. If  $o \in Objects$  and  $\sigma \in Types$ ,  $o \models \sigma$  means that  $o$  is of type  $\sigma$ .

We assume that an object can (at least) be described by its spatio-temporal properties, or, in other words, its spatio-temporal location. In that case, the classification “Jean  $\models \ll\lambda, t\gg$ ” expresses the fact the object “Jean” occupies position  $\lambda$  at time  $t$ .

A predicate (e.g. *to be tired, to walk*) is taken to denote an eventuality (Parsons, 1990)), and this is an *Object* in the model. As any other (abstract) *Object*, an eventuality can be described by a type (spatio-temporal location). For instance, we can specify at where and when someone has been walking:  $walk_x \models \ll\lambda, t\gg$

However, it is not possible to assign a spatio-temporal location to any kind of predicate. Individual-level predicates such as *to be blond, beautiful* ... (Carlson, 1977) cannot be described in this way. Nevertheless, note that when IL-predicates are transformed into tropes (e.g. *the blondness of Mary*) they can be described - in some particular cases - by a spatio-temporal location.

- **Constraint.** Constraints are strict entailments ((Lewis, 1973)). A strict entailment can be considered a standard entailment that has undergone the rule of necessitation.  $\neg\diamond\neg(p \rightarrow q)$  means that whenever  $p$  is true,  $q$  is also true. It is not possible that  $p$  is true but not  $q$  (as for standard entailment) nor that  $q$  is true but not  $p$  (differently from standard entailment).

The constraints are key to our interpretation of *avec*-instrument and manner. As we said, they are strict entailments, and, as such they express the notion of cause in counterfactual terms: it is not possible that  $p$  and  $\neg q$  or that  $\neg p$  and  $q$ .

Intuitively, for *avec*-instrument, they allow the expression that it is not possible that if John cuts the bread using a knife, then the knife does not have the necessary properties to cut the bread. Nor it is not possible that if the knife has the property to cut the bread, then John, using this knife, can not cut the bread<sup>6</sup>. This intuition also underlies the interpretation of *avec*-manner. Let “it is the case that John is watching TV and he is feeling pleasure” be a partial paraphrase for (3). *Avec* adds a constraint and the complete paraphrase becomes: it is not possible that, if John watches TV, then he does not feel pleasure nor that, if he feels pleasure, then he is not watching TV. It follows that it is not

necessary that John feel pleasure otherwise. Recall that whenever the property is not dependent on the eventuality denoted by the *VP*, the sentence is out. The model correctly predicts this fact.

A constraint between types is represented by a channel. A channel rests on infomorphisms, that we consider now.

(32) **Infomorphism.** An infomorphism is a pair of classifications  $(Object_1, Type_1, \models_1)$  and  $(Object_2, Type_2, \models_2)$  associated with two total functions  $f : Object_1 \rightarrow Object_2$  and  $g : Type_2 \rightarrow Type_1$  such that for  $o \in Object_1$  and  $\sigma \in Type_2 : f(Object_1) \models Type_2$  iff  $Object_1 \models g(Type_2)$

$$\begin{array}{ccc}
 Type_2 & \xrightarrow{g} & Type_1 \\
 \vdots & & \vdots \\
 \models_2 & & \models_1 \\
 \vdots & & \vdots \\
 Object_2 & \xleftarrow{f} & Object_1
 \end{array}$$

Let us consider an example:

(33) La fille avec un chapeau

*The girl with a hat*

(33')

$$\begin{array}{ccc}
 \ll \lambda_i \gg & \xrightarrow{g} & \ll \lambda_i \gg \\
 \vdots & & \vdots \\
 \models_2 & & \models_1 \\
 \vdots & & \vdots \\
 hat & \xleftarrow{f} & girl
 \end{array}$$

Let  $f(girl) = hat$ ;  $g(loc_i) = loc_i$ . Following definition (32) -  $f(Object_1) \models Type_2$  iff  $Object_1 \models g(Type_2)$ . By proper substitution we obtain:  $f(girl) \models loc_i$  iff  $girl \models g(loc_i)$  and then  $hat \models loc_i$  iff  $girl \models loc_i$ . This formula states that the hat is exactly the hat of the girl who wears it. Under this representation the localization of the hat depends on the localization of the girl who wears it. Infomorphism (33') links the entities in such a way that their spatio-temporal locations depend on each other.

Two infomorphisms sharing a common classification form a channel.

(34) **Channel.** A channel is a set of infomorphisms sharing a common classification called the *core* of the channel.

$$\begin{array}{ccccc}
 \alpha & \xrightarrow{f'} & f'(\alpha) \vdash g'(\beta) & \xleftarrow{g'} & \beta \\
 \vdots & & \vdots & & \vdots \\
 \models_1 & & \models_3 & & \models_2 \\
 \vdots & & \vdots & & \vdots \\
 Part_1 & \xleftarrow{f} & Whole & \xrightarrow{g} & Part_2
 \end{array}$$

The constraint  $f'(\alpha) \vdash g'(\beta)$  means that the association is not accidental (*validity of the inference*). It can be written, following ((Lewis, 1973)):  $\neg \diamond \neg (f'(\alpha) \rightarrow g'(\beta))$ .

Consider again example (33). We can state that the location of the girl determines the localization of the hat that she wears (33'). Another possible interpretation is that the girl and the hat form a **whole** “the girl with a hat” such that the girl on the one side and the hat on the other side are two parts of this whole (33’). This is obtained by instantiating scheme (34):

(33’)

$$\begin{array}{ccccc}
 \ll \lambda, t \gg & \xrightarrow{f'} & f'(\ll \lambda, t \gg) \vdash g'(\ll \lambda, t \gg) & \xleftarrow{g'} & \ll \lambda', t' \gg \\
 \vdots & & \vdots & & \vdots \\
 \models_{loc1} & & \models_{loc3} & & \models_{loc2} \\
 \vdots & & \vdots & & \vdots \\
 Part_1 : girl & \xleftarrow{f} & S: a girl with a hat & \xrightarrow{g} & Part_2 : hat
 \end{array}$$

Again, the constraint  $f'(\ll \lambda, t \gg) \vdash g'(\ll \lambda, t \gg)$  expresses a non-accidental linking between the position of the girl and the one of the hat she wears. Contrary to (33’) they are represented as two parts of a unique whole. This representation is particularly suitable for cases where there is an exact symmetry between the entities as for *A mum with her baby*  $\rightarrow$  *A baby with his mum*.

It is important to note that in a very abstract sense, a **situation** can be interpreted as a whole: it “keeps together” the entities that it involves. This implies, as in every part-whole relation, that the situation does not exist without its parts. More precisely, the *relations* among the entities create the situation.

The relation between two entities of the same situation (or whole) can be modeled by a channel. In this case, we can affirm that they are **coordinated** in the situation. By “coordination” we mean that:

- a. their properties depend on each other, and
- b. they co-participate in a unique action.

The arrows (representing functions) from the whole to the parts “extract” the participating entities; the arrows from the types of the parts to the type of the whole indicate that the descriptions of the parts are of a certain type **any**

**time**<sup>7</sup> the description of the situation is of a certain type. The types of the parts and of the whole can be the same as in this case.

In this way, our model can also integrate the notion of **situation-as-a-whole-dependence** of the coordination.

The idea that a situation can constitute the core of a Channel is fundamental to the definitions of instrumentality and manner and very likely extends beyond the scope of this study. Let us then develop it in the next section before we come back to *avec*.

### 3.1 The notion of situation type

The notion of situation is at the heart of the literature on situation semantics ((Barwise and Perry, 1983)) to which Channel Theory is strictly related. It is generally assumed that a situation is a structured part of the reality that the cognitive agent manages to pick out. Individuals, relations and spatio-temporal locations are the ingredients of situations.

Moreover, the cognitive agent is able to recognize situation types, that is to say, she is able to foresee how the entities will behave, given the knowledge that she has about the situation.

Let us consider two examples. In a situation where the agent observes some people in a queue, she will be able to foresee how the individuals will move, without having to observe the specific queue. A more complex situation consists of two entities related by a causal relation, for instance, a computer linked to a printer. Any time the agent makes a certain action on her computer (for instance she makes the request of printing a document by specific commands), the printer (linked to the computer that the agent uses) will print the document. The behavior of the printer is coordinated to that of the computer. Because she knows that the computer and the printer are linked to one another and that their behavior is coordinated, she knows that when she makes a certain action on the computer she will obtain another action from the printer.

In this respect, the situation where the computer is linked to a printer, i.e. “*computer*⊗*printer*” behaves as the abstract whole or the core of a channel. From now on, then, we will be considering situation types.

(35) *Situation type.* A situation type is a higher-order situation in which the behavior of the entities is predictable on the basis of their description.

Moreover, situation types are wholes that coordinate the behavior of the entities that they support.

(36) *Situation as whole.* A situation is the maximal entity supporting coordinated entities.

Of course, the very maximal entity is the universe. Following ((Devlin, 1991)) we will be considering only those situations that the cognitive agent can pick up in a limited spatio-temporal region.

Before we come back to *avec*, let us recall the main features of our model:

- it integrates tropes or spatio-temporal situated descriptions; namely the categorizations of the entities that depend on situation types;
- it integrates situations conceived as wholes keeping together the entities that they involve;
- it relies on constraints or causal relations linking the properties of the entities among each other and with respect to the situation.

## 4. A model for *avec*-instrument and manner

**4.0.1 An interpretation of *avec* in terms of channel.** We can now return to *avec*. Recall that our aim is to find the unique conceptual scheme which arises from these two meanings.

We have already emphasized that their similarities lie in the structuring relations among *X* and *Z* and the sub-events involving these entities. In particular, we have stated condition (30) that we repeat here for clarity reasons. Sub-event *e*<sub>3</sub> is particularly important: two entities *X* and *Z* are in a certain relation *with respect to the action denoted by the VP (situation dependence of the relation)*.

### (30) Sub-events: analysis of *avec*-instrument/manner and situation dependence

- *e*<sub>1</sub>: *X* performs the action denoted by the *VP*;
- *e*<sub>2</sub>: *Z* is causally linked (or non-accidentally related, or dependent on) to the action denoted by the *VP* (see ((Lewis, 1973)) for the notion of counterfactuality);
- *e*<sub>3</sub>: *X* and *Z* are in a certain relation with respect to the action denoted by the *VP (situation dependence of the relation)*.

We can now further abstract toward the common notion: *X* and *Z* are two *coordinated* entities with respect to two parameters given in (37):

(37) ***Avec*-instrument and manner and the abstract notion of coordination**

- a. they are coordinated with one another ( $X \otimes Z$ ):  $X$  controls - or is the source of -  $Z$ ;
- b. they are coordinated with respect to the main action:  $X$  and  $Z$  participate (in coordination) in the action denoted by the  $VP$  - causal or non-accidental link. -

It is possible to conclude that the properties of the entities are *regulated* or *coordinated* with respect to each other when the main action takes place. Conversely, the action can take place when the entities that it involves have such properties that they can enter in a relation of coordination. In other terms, there is a *causal constraint* between:

- a.  $X$  and  $Z$  and
- b. between the action on one side and ( $X \otimes Z$ ) on the other side.

As we have shown, these are features that Channel theory can easily express.

**4.0.2 Underspecified representation for *avec* in construction (1').**

We can now elaborate the unique, underspecified ((Pinkal, 1985), (Poesio, 1996)), possibly conceptual based-scheme for the meanings of *avec*-instrumentality and manner. In both of the cases the coordination has scope over the spatio-temporal locations of the entities involved in the situation. This means not only that the two entities share the same spatio-temporal location but also that it is *necessary* that, if one of them is in a certain location, the other be there too. This is so by virtue of the existence of a unique situation in which they are involved. We do not need to specify, at this point, the nature of  $X$  and  $Z$ .

(38) Underspecified representation for *avec*-instrument and manner

$$\begin{array}{ccccccc}
 \ll \lambda, t \gg & \xrightarrow{f'} & f'(\ll \lambda, t \gg) \vdash g'(\ll \lambda, t \gg) & \xleftarrow{g'} & \ll \lambda', t' \gg \\
 \vdots & & \vdots & & \vdots \\
 \models_{loc1} & & \models_{loc3} & & \models_{loc2} \\
 \vdots & & \vdots & & \vdots \\
 X & \xleftarrow{f} & S & \xrightarrow{g} & Z
 \end{array}$$

This representation can be paraphrased in the following way: entity  $X$  and entity  $Z$  are in a certain spatio-temporal location because the situation  $S$  takes place in this very same spatio-temporal location. Reciprocally, the situation  $S$  takes place in a certain spatio-temporal location because there is a link between the spatio-temporal locations of the entities it supports. This link, as it is expressed by the constraint  $f'(\ll \lambda, t \gg) \vdash g'(\ll \lambda, t \gg)$ , is not accidental.

As expected from (37),  $X$  and  $Z$  are coordinated with one another and with respect to the main action.

## 4.1 Representation of *avec*-instrument and manner

Representation (38) is differently instantiated by *avec*-instrument and manner.

In the case of *avec*-instrument,  $X$  *uses* or generally *controls*  $Z$ , in the case of *avec*-manner,  $X$  is the psychological source of  $Z$ .

Only the coordination of the spatio-temporal locations of the parts has to be represented at the semantic level. In this case, the “parts” of the situation, are, on the one side, the action involving the actor/agent ( $NP_1 VP$ ), and, on the other, the property/action involving the entity denoted by  $NP_3$  ( $NP_1 VP_{interpolated}$ ). These two are, respectively, an overt and an interpolated predicate, and they are considered as *Objects* in the model. Recall, in fact, that *Objects* stands for any kind of abstract entity: properties or eventualities. *Types* describe them, assigning, minimally a spatio-temporal location.

It is important to note that control or psychological source relations are not introduced as types. Instead, they are treated as abstract relations between properties of singular entities. More precisely, they are treated as holding between the property that the agent has (i.e. the action she is involved in ( $NP_1 VP$ )) and the property/action involving the entity denoted by  $NP_3$  ( $NP_1 VP_{interpolated}$ ). As we have just noted, these correspond to two *Objects* in the model, or to two “parts” of the situation described by the sentence.

Let us illustrate this by considering, in turn, *avec*-instrument and -manner.

## 4.2 Representation of *avec*-instrument

In the case of *avec*-instrument there exists a situation in which the property of the agent effectuating a certain action is coordinated with the property of a certain entity.

Let us consider the situation described by the sentence

(39) Jean a brûlé le tapis avec une cigarette  
*John burned the carpet with a cigarette*

In this case, the *Objects* will be *burningJohn* / *burningcigarette*. These are properties of John and the cigarette. They are treated as *Objects* (see section 3., (31)) and they are assigned a *Type*. In particular, they are described by their spatio-temporal location, specifying at what place and time John and the cigarette have the property of burning the carpet.

These spatio-temporal locations are coordinated. It follows that John burns the carpet with the same cigarette that that burns the carpet in the same spatio-temporal location.

Given the knowledge that the speaker has about the relation between an individual and a cigarette, she can conclude that John was smoking or manipulating the cigarette in some way, i.e. he was “controlling” or “using” it. However, this is lexical or contextual information that is not encoded in the semantics of *avec*-instrument.

Representation (38a) is then a possible instantiation of (38) for (39).

(38a) Representation of *avec*-instrument

$$\begin{array}{ccccccc}
 \ll \lambda, t \gg & \xrightarrow{f'} & f'(\ll \lambda, t \gg) \vdash g'(\ll \lambda, t \gg) & \xleftarrow{g'} & \ll \lambda', t' \gg \\
 \vdots & & \vdots & & \vdots \\
 \models_{loc1} & & \models_{loc3} & & \models_{loc2} \\
 \vdots & & \vdots & & \vdots \\
 \text{burning}_{John} & \xleftarrow{f} & S & \xrightarrow{g} & \text{burning}_{cigarette}
 \end{array}$$

This representation can be paraphrased in the following way: in the situation where John is burning the carpet with a cigarette, the spatio-temporal locations of the property of John burning the carpet and of the property of the cigarette burning the carpet are related. It follows that John is burning the carpet with the cigarette that is burning the carpet. The agent can *infer* that John is smoking or using/playing with the cigarette in some way.

Note that lexical or contextual factors make explicit what property of the entities, and of *Object*<sub>2</sub> in particular, are called into play. In this case, the reconstruction is straightforward. In some other cases, it can be more complex and totally context dependent. Almost anything can be used as an instrument and the relevant characteristics that are expected to cause the action are very high in a given context (see section 2.1.2).

### 4.3 Representation of *avec*-manner

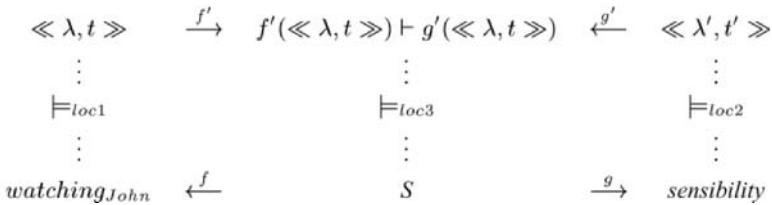
*Avec*-manner is the other possible instantiation of (38). In this case what makes the situation exist is that the agent is the psychological source of the feeling. This “psychological-source” relation, again, is not a type. Consider the following sentence.

(40) Jean regarde la télévision avec intelligence  
*John is watching TV wisely*

As for *avec*-instrument, the spatio-temporal location of the property of watching TV of John is coordinated with the spatio-temporal location of being

wise. The coordination of the spatio-temporal location of a certain property of John (*regarder/watching*) with the spatio-temporal location of the wisdom of John allows the expression that John is not necessarily wise otherwise. According to what the sentence tells us about John, only his watching of TV is wise.

(38b) Representation of *avec*-manner



This representation can be paraphrased in the following way: in the situation where John watches TV, John is watching TV and he is sensible. It follows that the observer can only know that John's TV watching is wise, but not necessarily that John is wise otherwise.

## 5. Conclusion

In this paper we have tried to provide a formal account for the well-founded intuition that instrument and manner are two related notions. We have argued that it is possible to provide an explanation considering their lexicalizations.

*Avec* has, among others, the meanings of instrument and manner. We have proceeded by a bottom-up analysis individuating first the descriptive parameters and then their common conceptual ground.

We have shown that *avec*-manner and instrument share some essential characteristics:

- the properties of the denotation of the head noun of the  $NP_3$  have to be situation-dependent and causally related to the action described by the *VP*.
- both of these meanings can be analyzed in terms of the sub-events involving *X* and *Z*. These entities are coordinated (depend on each other) and both, by virtue of their coordination, participate in the action described by the *VP*.
- The different natures of the relation between *X* and *Z* distinguish the two meanings. These are “control”, or “psychological source” relations. They depend on lexical or contextual information and are not encoded in the semantic representation of *avec*-instrument and *avec*-manner.

The notion of causality we have been referring to is strictly related to that of counterfactuality ((Lewis, 1973)) and amounts to a non-accidental relation among objects via their properties or, more generally, to a relation of dependence.

We have proposed a model inspired by the Channel Theory of ((Barwise and Seligman, 1997)) and we have emphasized two points in particular:

- the model uses types and thus takes into account the categorization that human beings make out of the entities in a given situation or state of affairs,
- it represents the coordination of the properties of two entities (dependence of the descriptions) within a unique situation that keeps them together.

We have then interpreted the characteristics common to *avec*-instrument and -manner along the lines of these formal features. A coordination of *X* and *Z* in the situation enhanced by the *VP* is minimally required. We have expressed the underspecified coordination of *X* and *Z* by the non-accidental linking of their spatio-temporal locations. This coordination is represented by the channel.

We can conclude that the intuition according to which there is a link between the notions of manner and instrument is well founded, but that their similarity requires a high degree of abstraction to be captured.

An explanation based on the observation that the types of the head noun of the *NP*<sub>3</sub> are not the same is too simple. Moreover, no distributional criteria have been found for *avec*-manner, such that they can clearly delimit the set of acceptable nouns. In the same way, because almost anything can be used as an instrument, it is impossible to individuate specific classes of nouns denoting instruments.

The last open question is to know whether the similarities between the notions of instrumentality and manner are lexically driven or if they are more general, possibly universal.

## 5.1 How universal is the relation between the notions of instrumentality and of manner?

At this point we can conclude that instrumentality and manner are two related notions that share an underspecified mental representation.

The study we have presented here could lead to the conclusion that the similarities between these two notions are lexically driven: *avec* (*with*) is the only preposition that instantiates both of them. It would follow that the notion of dependence (in channel-theoretic terms) is lexically driven. This conclusion seems to be confirmed by a complete study of the meanings of *avec* ((Mari, 2003)). Here it is shown that this preposition is specialized in the instantiation

of the notion of “association as dependence”. This is specified in two ways: association as influence and association as spatio-temporal link. Instrument and manner belong to this second class.

This might not be the final conclusion, though. In fact, it does not follow that instrumentality and manner are completely unrelated in other cases. There are certainly some *resemblances* ((Wittgenstein, 1953)) that go beyond the lexical meaning of *avec* and cognitively relate these two notions. To show this, one has to consider, separately, other lexicalizations of instrumentality and manner. This goes beyond the scope of this paper, but independent studies on instrumentality on one side ((Mari and Saint-Dizier, 2003)) and on manner adverbs on the other ((Molinier, 1984)), seem to confirm this hypothesis. The meaning of instrumentality and manner can be explained in many cases by a causal connection between the entities involved in the action denoted by the *VP*. This connection can take the form of a control or a psychological-source relation and can be represented as a coordination of descriptions.

Moreover, with respect to the model, the notions of situation type and constraint are general enough to lead us to believe that we have reached a fundamental point of similarity. Again, one could argue that that *avec* is specialized in instantiating these notions in language. It is known nevertheless that other items and constructions behave, conceptually, in a similar way ((Jayez and Mari, 2004)). This is why we can risk affirming that the similarities between instrumentality and manner formulated in terms of causal relation are only lexically driven by *avec* but generate at a higher level of abstraction, involving a notion of causality as non-accidentality. The next step of our analysis will be to compare *avec* and the notions it expresses with other items whose senses can be formulated in terms of coordination and constraints on descriptions. Meanwhile, we can add *avec* to the list of items that express causality in language<sup>8</sup>, even if it captures a more abstract aspect of this notion than the other items already admitted to this list.

## Notes

1. *X, Y, Z* are variables for objects or entities, including abstract objects and events
2. It is important to note that *avec* does not solve the ambiguity agentive/non agentive interpretation of the main predicate (e.g. *John s'est brûlé avec de l'huile bouillante / John burned himself with boiling oil*). It only requires that *X* controls *Z* such that this control has consequences on the action. Whether the action is brought about voluntarily or not, it is not a question that is related with the semantics of *avec*.
3. Note that the acceptabilities can vary from one language to another.
4. It is obviously the case that “being marmalade pure sugar” entails “being marmalade”. This is however not an issue, here. In accordance with Anscombe (*ibid.*) what is at stake here, is the ontological link existing between the substance and its property, in a topology where we are focusing the properties of “marmalades pure sugar”
5. An eventuality ((Parsons, 1990)) is any kind of temporal entity, static or dynamic (see ((Binnik, 1991)) for an introduction).
6. The hearer assumes that John has the ability to cut the bread if he uses a knife that allows him to do so and that there are no other obstacles.

7. Validity of the inference
8. See, for a survey, ((Nazarenko, 2000))

## References

Anscombe, J-C. (1990). "Article zéro et structuration d'événements." In M. Charolles, S. Ficher, J. Jayez (eds.), *Le discours, représentations et interprétations*. Nancy : Presses Universitaires de Nancy, pp. 265-305.

Barwise, J., Perry, J. (1983). *Situations and Attitudes*. Cambridge MA: MIT Press.

Barwise, J., Seligman, J. (1997), *Information Flow*. Cambridge: Cambridge University Press.

Binnik, R.I. (1991). *Time and the verb. A Guide to Tense and Aspect*. Oxford: Oxford University Press.

Cadiot, P. (1997). *Les prépositions abstraites en français*. Paris: Armand Colin.

Carlson, G-N. (1977). *Reference to Kinds in English*. Thèse de l'Université de Massachusetts.

Corblin, F. (1995). *Les formes de reprise dans le discours*. Rennes : PUR.

Devlin, K. (1991). *Logic and Information*. Cambridge: Cambridge University Press.

Flaux, N., Van de Velde, D. (2000). *Les noms en français, esquisse de classement*. Paris : Ophrys.

Jayez, J., Mari, A. (2004). "Togetherness". *Sinn und Bedeutung* 9: Nijmegen.

Godard, D., Jayez, J. (1993). "Towards a Proper Treatment of Coercion Phenomena." *Proceedings of the 6th Conference of the European Chapter of the ACL*, Utrecht, pp. 168-177.

Leeman, D. (1995). "Pourquoi peut-on dire *Max est en colère* mais non pas \**Max est en peur* ? Hypothèses sur la construction *être en N*". *Langue Française* 105, pp. 55-69.

Lewis, D.K. (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Mari, A. (2002). "Under-specification and contextual variability of abstract prepositions : a case study." *Proceedings of ACL-SIGSEM Workshop*: Philadelphia, pp. 17-24.

Mari, A. (2003). *Principes de catégorisation et d'identificatioin du sens. Le cas de "avec" ou l'association par les canaux*. Paris: L'Harmattan.

Mari, A., Saint-Dizier, P. (2003). "Conceptual Semantics for Prepositions denoting Instrumentality." *Proceedings of GL03*: Geneva, pp. 222-229.

Molinier, Ch. (1984). *Etude syntaxique et sémantique des adverbes de manière en -ment*. Thèse de Doctorat, Université de Toulouse-le-Mirail.

Nazarenko, A. (2000). *La cause et son expression en français*. Paris: Ophrys.

Parsons, T. (1990). *Events in the Semantics of English*. Cambridge MA: MIT Press.

Pinkal, M. (1985). *Logic and Lexicon*. Oxford: Oxford University Press.

Poesio, M. (1996). "Semantic Ambiguity and Perceived Ambiguity". In K. van Deemter and S. Peters, (eds.), *Semantic Ambiguity and Underspecification*. Stanford : CSLI Lecture Notes. pp. 159-201.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge MA: MIT Press.

Simons, P. (1994). "Particulars in Particular Clothing: Three Trope Theories of Substance." *Philosophical and Phenomenological Research* 54, pp. 507-575.

Spang-Hanssen, E. (1963). *Les prépositions incolores du français moderne*. Copenhague: G.E.C. Gads Forlag.

Talmy, L. (1976), "Semantic Causative Types." In M. Shibatani (ed.), *Syntax and Semantics 6: The Grammar of Causative Constructions*. New York: Academic Press, pp. 43-116.

Wierzbicka, A. (1996). *Semantics Primes and Universals*. Oxford: Oxford University Press.

Wittgenstein, L. (1953). *Investigations philosophiques*. Trad. P. Klossowski. Paris: Gallimard.

## Chapter 18

# A CONCEPTUAL SEMANTICS FOR PREPOSITIONS DENOTING INSTRUMENTALITY

Alda Mari

*ENST - CNRS*

*46, rue Barrault - 75013 Paris - France*

*mari@enst.fr*

Patrick Saint-Dizier

*IRIT - CNRS*

*118 Route de Narbonne, 31062 Toulouse, France*

*stdizier@irit.fr*

**Abstract** In this paper, we present a semantic analysis and a representation for prepositions dealing with instrumentality. The abstract parameters defining instrumentality are elaborated and a model of the interactions agent-object-instrument is proposed and implemented using the Lexical Conceptual Structure.

**Keywords:** Instruments, Lexical Conceptual Structure.

### 1. An analysis of the primitive notion of instrumentality and its lexicalizations

The notion of instrument seems to appear at a relatively early stage of the semantico-cognitive development of children and has often been considered as a primitive notion (Wierzbicka, 1992). Moreover, at first glance, its contents can be expressed by a very intuitive paraphrase: an instrument is an object used to realize an action or to reach a certain goal. However, behind this apparent simplicity, the parameters defining this notion from a semantic and lexical point of view are complex and subtle. Our study of this phenomenon is both analytical and formal: it considers the abstract notion as well as its possible lex-

icalizations. It aims at providing a comprehensive analysis for the observable meanings while addressing the question of the existence of an underspecified representation for the notion of instrumentality.

The notion of instrumentality, in spite of its importance in language and thought, has not received much attention besides the conceptually oriented work of (Talmy, 1976). (Mari, 2003) and (Mari, this volume) is one of the first attempts to analyze its semantics and to provide a logical account.

This work concentrates on four prepositions, representative of the notion of instrumentality in French: *par* (approximately ‘by’), *grâce à* (approximately ‘thanks to’), *au moyen de* (approximately ‘by means of’), and *avec* (approximately ‘with’). A number of uses of these four prepositions have been collected in various corpora. Our analysis emerges from these uses; for illustrative purposes, a prototypical example has been selected for each preposition use.

This chapter is organized in three parts. The first part (section 2) attempts at identifying the abstract parameters that define the notion of instrumentality, constraints on its lexicalisations and on its contextual values. In the second part (section 3), we suggest a model using the Lexical Conceptual Structure (LCS) (Jackendoff, 1990) and elements of the Generative Lexicon (GL) (Pustejovsky, 1995) that we settle within a compositional framework. Our general aim is to provide underspecified representations for the abstract notion, based on the LCS paired with conditions and contextual values for each type of use. A typed  $\lambda$ -calculus for computing meanings is also provided. In the third part of the paper (section 4), we come back to the prepositions studied in the first part, and we show how our model allows us to implement their semantics. We also present an underspecified representation for the notion of instrumentality, which can possibly be used as a template for the analysis of prepositions in other languages and we conclude in section 5.

In order to settle our analysis, let us begin by stating the difference between abstract forms and meanings in contexts, and by briefly discussing some methodological issues.

## 1.1 Three levels of analysis: notions, senses and values

In the semantic literature, there exist two ways to capture the notion of instrumentality: the first one considers the object type expected by the preposition complement (Poncet-Montagne, 1991); the second one consists in identifying the possible relations between the sub-event denoted by the *VP* and the “causing” sub-event, i.e. the one involving both the subject and the preposition complement: for *Jean mange avec une cuillière* (*John eats with a spoon*) the causing sub-event is “Jean uses a spoon”. Talmy’s work on force relations (Talmy, 1976) relies on this second option. Nevertheless, considering only the

cognitive universal aspect of this notion, his work ignores the constraints on the structure of these sub-events, the specificities related to their possible lexicalizations and the parameters defining the control relations among the entities involved within these sub-events.

Our analysis, based on event structure, integrates these parameters. Our contribution extends towards the interpretation of the complex relations among sub-events and the entities that they involve, with respect to three levels of representation. Given the dichotomy between notions (or senses) and values (or meanings) (Mari, 2000) that we understand as a difference between semantic category and contextual instantiations (Poesio, 1996), we consider instrumentality at the following three levels of abstraction (terms are borrowed from Poesio):

- 1 the underspecified form of the notion of instrumentality or cognitive category,
- 2 the underspecified forms of the possible lexicalizations of the notion: definition and representations of the sense(s) of *par* (*by*), *grâce à* (*thank to*), *au moyen de* (*by means of*), *avec* (*with*),
- 3 the contextual values of the sense(s) of these prepositions.

The final aim of our study is twofold: to propose a formal implementation of the contextual values of each of the prepositions, and to propose a representation for the underspecified form of the notion of instrumentality.

In order to reach the abstract notion, we proceed by a bottom-up analysis and concentrate, in the next section, on the contextual values of the prepositions.

## 2. Analysing the notion of instrumentality via its lexicalizations

In this section we present our analysis of the meanings of the four prepositions *par* (*by*), *grâce à* (*thanks to*), *au moyen de* (*by means of*), *avec* (*with*). One of the major results is that the semantics of these prepositions have scope, in the proposition, over the verb predicate. This is implemented in our framework by considering three sub-events involving the agent the instrument and the main action, and defining precise scope relations between them.

The presentation of our analysis of the prepositions denoting instrumentality is organized as follows: we first define the event structure related to the notion of instrumentality (section 2.1), and then present the essential features of the data defining the senses of the four prepositions at an informal level (section 2.2), focusing on the parameters that serve the instantiation of this notion in order to achieve a deeper and computationally tractable formal model. Before we go on with the presentation of the model, in section 2.3, let us present a

brief note on the interactions between the semantics of the prepositions and the semantics of the main predicate, where we claim that the first has wider scope over the latter.

Let us note that prepositions are language-specific and that our analysis initially applies to French prepositions. The methodology and, possibly the results, can be extended to other languages and to other prepositions, not necessarily those mentioned in the glosses, though.

## 2.1 The event structure of the notion of instrumentality

To see how the notion of instrumentality can be decomposed into three sub-events involving entities having complex relations between each other, let us consider the following example :

*John cuts the bread = e3*

*John uses a knife = e2*

*The knife has the ability to cut the bread = e1.*

The syntactic surface structure is:

*NP0 V NP1 PrepInstr NP2*

In the following discussion, *I*, *J* and *K* represent the denotations of *NP0*, *NP1* and *NP2*. Since there are many syntactic structures that accept an object NP (*NP1*) we could alternatively have the more generic structure: *NP0 - VP - PrepInstr - NP2*. However, since we need to take into account the *NP1* in a number of cases, we leave it explicit. Our analysis assumes the existence of the following sub-events and entities:

- 1 The sub-event (*e<sub>1</sub>*) implying the instrument (*K*) and the action (*VNP<sub>1</sub>*).
- 2 The sub-event (*e<sub>2</sub>*) implying the actor / agent (*I*) and the instrument (*K*).
- 3 The sub-event (*e<sub>3</sub>*) implying the actor / agent (*I*) and the action (*VNP<sub>1</sub>*).

The formula (i) makes explicit the relations existing among these sub-events:

(i)  $(e2(e1)) \Rightarrow e3$

It expresses the fact that “because the knife has the ability to cut the bread (*e1*) and that John uses the knife (on the bread) (*e2*), then John cuts the bread (*e3*)”. The application of *e2* on *e1* entails *e3*.

It is clear that the preposition has scope over the main predicate. The main motivation for this choice is that the relation between the agent and the instrument on the one hand and the instrument and the main action on the other hand, determines the relation between the agent and the action itself and the way it is carried out.

## 2.2 The data: informal definitions, notions of sub-event and control relations

The general instrumentality schema is instantiated differently depending on the lexicalizations of the notion of instrumentality used. Methodologically, we begin by considering these lexicalizations to abstract later over the cognitive notion and representation of instrumentality. For each of the prepositions, we present a typical example taken from our corpus, an informal definition, and a sub-event based paraphrase in the lines of (i).

### 2.2.1 Par (by). Typical exemple:

(1) *Les alpinistes ont atteint le sommet par ce chemin / The alpinists reached the top by this trail*

Informal definition: “*e1 = K*, is in a certain disposition or state such that it can have a certain effect; *e2 = it does an action on K* which is of the type of *e1* or that entails *e1*; *e3* is obtained”.

Paraphrase in terms of sub-events (1): “the trail has the property of reaching the top of the mountain (*e1*), the alpinists walk on this trail (*e2*) and so, they reach the top of the mountain (*e3*)”.

### 2.2.2 Grâce à (thanks to). Typical example:

(2) *Le tourisme prospère grâce au Canal du Midi / Tourism thrives thanks to the Canal du Midi*

Informal definition: “*e1 = the instrument (J)* has certain properties (that need to have a priori positive consequences); *e2 = the actor (I)* is positively influenced by these properties; *e3* the actor *I* benefits from *J*”.

Paraphrase in terms of sub-events (2): “the Canal du Midi is a touristy attraction (*e1*), tourism benefits from the presence of the Canal du Midi (*e2*) and thus tourism thrives (thanks to the Canal du Midi) (*e3*)”.

### 2.2.3 Au moyen de (by means of). Typical exemple:

(3) *Il s'est brûlé au moyen d'huile chaude / He burned himself by means of boiling oil*

(4) *Ils ont ouvert la porte au moyen d'un cric / They opened the door by means of a jack*

Informal definition: “*e1 = the instrument (K)* is such that it can perform an action; *e2 = the agent (I)* controls the action that the instrument can perform; *e3 = the agent performs the action*”.

Paraphrase in terms of sub-events (3): “a boiling oil can burn (*e1*), John uses boiling oil to burn himself (*e2*) and thus he burns himself (*e3*)”.

### 2.2.4 Avec (with). Typical exemple:

(5) *Jean s'est brûlé avec de l'huile chaude / John burned himself with boiling oil*

Informal definition: “*e1* = the instrument is such that it can perform an action; *e2* = the actor (or agent) controls the instrument without controlling the action that it can perform; *e3* = the agent uses the instrument and the action is realized by the instrument”.

Let us notice that *avec* has two possible interpretations: either *I* is an actor (in this case John unwillingly burns himself), or an agent (John is willing to burn himself). In this second case *avec* is a synonym of *au moyen de*. In the remainder of this paper we consider the first interpretation only. Paraphrase in terms of sub-events for (5) are: “boiling oil can burn (*e1*), John uses the boiling oil (*e2*) and he burns himself (*e3*)”.

### 2.3 A note on the interaction between the semantics of the preposition and the main predicate

One of the objections to our analysis, is that the ambiguity intentionality vs. non-intentionality is already present in the semantics of the predicate. *Burn* can have an intentional or a non intentional interpretation which is not due to the preposition.

Let us emphasize that we are not claiming that the preposition is ambiguous, or that there are two different prepositions *avec* denoting instrumentality. The predicate is ambiguous, and the preposition can possibly solve the ambiguity or not. In (3), *au moyen de*, solves the ambiguity, whereas in (5), *avec* does not.

The task of the preposition is twofold: on the one hand it specifies in what particular way the objects that it selects are involved in the main action, on the other hand it specifies a particular relation between the agent / actor and the instrument. As a result of these two constraints, it may or may not disambiguate the predicate, in cases where it is itself ambiguous. Note that, if one of the meanings of the predicate is not compatible with one of the constraints that the preposition imposes, the sentence cannot be accepted. Consider, for instance, the impossibility of *\*John est apprécié au moyen de sa beauté* (*John is appreciated by means of his beauty*), where the agentive relation between *I* and *K* is incompatible with the non-agentive relation between *I* and the event denoted by the *VP*.

This view of the role of the prepositions also explains why they may be considered to have wider scope over the predicate, at least in a semantic perspective. The preposition, introducing an entity into the scene, also specifies its relations to the other entities, the actor / agent and the action. This, as we have noted it, affects the way in which the actor / agent carries the action at stake. In this respect, we can say that the semantics of the main predicate is a func-

tion of the semantics of the preposition, and the new relation that it establishes between the subject and the action in which it is involved.

Two important issues remain open, among many others. The first one is to know whether the observation that the semantics of the prepositions influences that of the main predicate is general and can be extended to any kind of preposition (abstract - e.g. *à (to)*, *de (of)* - the so-called mixed - e.g. *pour (for)* - and the concrete ones - e.g. spatial prepositions<sup>1</sup>). The second one is to know whether the classical distinction between argument and adjunct can influence the way the semantics of prepositions interacts with that of the main predicate. These are general questions that are outside the scope of this paper (see (Bonami, 1999)).

### 3. The logical model

In this section we present the logical model that we will be using to implement preposition senses. After a short presentation of the main principles of the Lexical Conceptual Structure, we introduce the additional primitives necessary to adequately encode the notion of instrumentality. We then provide an example for each preposition and show how we can abstract over these prepositions to get a generic representation for instrumentality.

#### 3.1 Main Principles of the Lexical Conceptual Structure (LCS)

The LCS owes much to the former Lexical Semantics Templates. It gained its popularity via Jackendoff's improvements. The LCS is mainly organized around the notion of movement (change of location), the other conceptual fields being derived by analogy, in a more or less natural way. We consider the LCS as a semantic representation language and as a methodology for describing the semantics of predicative forms. It is indeed clear that the primitives it is composed of are not exhaustive enough. Moreover, these primitives remain abstract and can be viewed as macros: in a language processing system, they may need to be interpreted, e.g. in terms of Euclidean geometry for primitives dedicated to the expression of localization.

The LCS language is composed of three elements: conceptual categories, also called parts of speech: *thing*, *event*, *state*, *place*, *path*, *property*, *purpose*, *manner*, *amount* and *time*. These are used to type the different LCS structures. The LCS also has a number of conceptual primitives. The most important ones cover the notions of *change* (GO), *state* (BE), and *cause* (CAUSE). Lower level primitives mainly include primitives encoding prepositions: FROM, TO, AT, ON, etc. In our framework, we consider that we need 64 such primitives (Cannesson et ali. 01). Finally, the LCS has semantic fields: *loc*, *temp*, *poss*, *epist*, *comm*, etc. designed to specialize the above set of primitives to a certain field:

$GO_{+loc}$  is a change in the localization domain, while  $GO_{+poss}$  is a change of possession.

LCS forms can be read quite easily, for example, the verb run is represented as follows:

$[\text{event } CAUSE([\text{thing } I],$   
 $[\text{event } GO_{+loc}([\text{thing } I, I [\text{path}])])]$

which can be paraphrased as: I (the subject, and only argument) is the cause of an event which is a change of localization ( $GO_{+loc}$ ) of itself along a certain path which is left underspecified (possibly instantiated by a PP). The LCS is particularly appropriate for representing information of a predicative nature, it needs to be paired with other types of representation paradigms (e.g. attribute-value pairs, scales, etc.) for the non-predicative information. A number of abstract verbs (as common as *write* or *read*) are quite difficult to represent in LCS in a 'natural' way.

In the representations given below, the LCS is paired with a typed  $\lambda$ -calculus and underspecification, allowing for (1) the introduction of information coming from arguments or from inferences and (2) the implementation of the principle of compositionality. The above example can then be rewritten as follows:

$\lambda I, \lambda P : [\text{path}]$

$[\text{event } CAUSE([\text{thing } I],$   
 $[\text{event } GO_{+loc}([\text{thing } I], P)])]$

where P is constrained to be any structure of type path.

## 3.2 Modeling the actor-agent / action / instrument relations

Let us now model the relations among the event *ei* presented above. For that purpose, we need to introduce two sets of primitives to characterize (1) the different levels of control that the actor / agent (*I*) has on the instrument (*K*) and (2) the degree of commitment of the instrument in the action. For example contrast *cut the meat with a knife* with *eat soup with a spoon*: in the first case the knife does the cutting whereas in the second, the spoon is just used as a tool that facilitates the action, it does not do the eating.

**3.2.1 The actor-agent (*I*) / instrument (*K*) relation: (e2) .** The control that the actor / agent has on the instrument varies considerably and can be expressed by means of three different primitives in the LCS:

- UNDERGO: the actor has no control on the instrument or on its properties.
- SELECT: the agent uses the instrument and has some control on it. Nevertheless, while doing a certain action with the instrument, the actor does not necessarily plans to do the action denoted by *VNP1*.

- CONTROL: the agent controls the instrument, in order to realize the action denoted by  $VNP1$ .

**3.2.2 The instrument ( $K$ ) / action ( $V\,PN\,I$ ) relation: ( $e1$ ) .** According to the commitment of the instrument in the action being performed, this relation can also be instantiated by three different primitives in the LCS:

- BE: the instrument has some intrinsic properties such that even being passive, it necessarily participates to the action denoted by  $VNP1$ .
- REACT: the instrument, while being controlled and activated by the agent with respect to a particular property, participates to the action denoted by the  $VNP1$  via another property, unexpected and uncontrolled by the agent.
- ACT: the instrument has an intrinsic property that contributes, via the agent, to the success of the action. The primitive ACT, contrary to the primitive BE, expresses the fact that the instrument is not passive, but that it participates to the action.

The relation  $e3$  does not need any additional primitive to be adequately represented.

### 3.3 Preposition senses and primitives: towards a formal definition

Before we proceed to the modeling of the four prepositions presented in section 2, let us summarize the way each of the primitives that we have introduced can be associated with a preposition. For each preposition, the notion of instrumentality emerges as a combination of relations among the entities denoted by the  $NPs$  and the action described by the verb:

relation $K$ to $VP$ ( $e1$ )	Relation $X$ to $Z$ ( $e2$ )		
	UNDERGO	SELECT	CONTROL
BE	<i>grace a</i>	<i>par</i>	
REACT		<i>avec</i>	
ACT			<i>avec, au moyen de</i>

It is now clear that the contents of  $e3$  is the result of the interaction between on the one hand the way the instrument is involved in the main action, and, on the other hand the relation between the subject and the instrument. Introducing  $e3$  and specifying its relations with the subject and the action, allows us to define the way the subject is involved in the action.

Consider now the examples above encoded by means of these primitives. We leave apart the main action, namely the event involving  $I/VP$  (i.e.  $e3$ ), to

concentrate here on  $e1$  and  $e2$ . We recall that the relation between these two events is the temporal inclusion:  $(e2 \ (e1))$ .

**Grâce à:** *Le tourisme (I) prospère (VP) grâce au Canal du Midi (K) / Tourism thrives thanks to the Canal du Midi*

From the above set of primitives, this statement can be represented schematically as follows:

$UNDERGO(I, K) \wedge (prop(K) \Rightarrow (IVP))$

Tourism benefits from the fact that the Midi Canal is a nice canal (noted here as  $prop(K)$ , a function that extracts a relevant property of  $K$ ). The instrument has an effect through its own properties.

**Par:** *Les alpinistes (I) ont atteint le sommet (VP) par ce chemin (K) / The alpinists reached the top by this trail*

A possible definition in terms of primitives is:

$SELECT(I, prop(K)) \Rightarrow (IVP)$

The alpinists choose the trail for some of its properties, without controlling it. They are nevertheless active in the choice of a particular trail, probably because they know that it will enable them to reach the top of the mountain.

**Avec:** *Jean (I) s'est brûlé (VP) avec de l'huile chaude (K) / John burned himself with boiling oil*

In terms of the above primitives, we have:

$SELECT(I, K) \wedge REACT(property(K)) \Rightarrow (IVP)$

John uses boiling oil that has the property of burning, without willing to burn himself. The property of burning is activated by REACT. The consequence is that John burns himself accidentally.

**Au moyen de:** *Jean (I) s'est brûlé (VP) au moyen d'huile chaude (K) / John burned himself by means of boiling oil*

We have in this case:

$CONTROL(I, K) \wedge ACT(property(K)) \Rightarrow (IVP)$

John uses the oil for its burning properties in order to deliberately burn himself. We strengthen the relation between  $K$  and the  $VP$  to indicate that it is a particular property of the oil that is looked for by the subject in order to achieve his aim.

## 4. LCS representation of preposition senses and instances

In this section, we first show (section 4.1) how the meaning of these four preposition senses can be represented and then (section 4.2) consider the abstract underspecified representation for the notion of instrumentality, that can be factored out from the representations of the contextual values.

## 4.1 Representation of the contextual values

The LCS provided for the four prepositions has a very regular structure that reflects the sub-event construction typical of instrumentality: the first part of the general form associated with the sense of the preposition describes the nature of the control of the subject *I* on the instrument *K* (*e2*), the second part accounts for the properties of the instrument (*K*) that are useful for the action to be realized (*e1*), while the third part describes the action itself (*e3*). As presented above, *e2* has wider scope over *e1*. Note that the theme *J* is only present in the verb representation within *e3*.

It is important to note that, in the calculus given below, *PPs* are generally analyzed as propositional adjuncts: their representation embeds the verb representation and not the reverse as it is in general the case for predicate arguments, which are included into the verb's representation. Syntactic alternations provide a strong argument in favor of this interpretation: these constructions generally undergo the Possessor-Subject (transitive) alternation (2.13.4, Levin, 1993), also valid for French (Saint-Dizier, 1998), which clearly indicates that the PP has wider scope over the whole proposition.

In all of the following LCS, let *I*, *J*, *K* be the variables representing respectively *NP0*, *NP1* and *NP2*; let *T* be the ontological type of the verb *VERB* of the proposition. For each of the prepositions, we give the general semantic form in LCS and then the representation of the selected typical example. Additional operators and considerations are introduced below when used.

### 4.1.1 Par. General form:

$$\lambda I, \lambda K, \lambda J,$$

$$[\text{event } CAUSE([\text{event } SELECT([I], [\text{state } BE_{+T}([K],$$

$$AT_{+T}([\text{telic-of}(K, J)]))])],$$

$$[\text{event } BECOME([I], [\text{event } GO_{+T}([I],$$

$$[\text{path } AT_{+T}([\text{VERB}([I], [J])])])])])]$$

The function *telic-of*(*K*, *J*) extracts in the telic role of the noun *K* a predicate whose argument types are subsumed respectively by the types of *K* and *J*. Telic role is a direct reference to the Generative Lexicon (Pustejovsky, 95), where predicates in the telic role represent actions or uses, in general prototypical, of the object *K*.

The primitive *BECOME* characterizes accomplishments. It emphasizes in general the state resulting from the action described by the verb. In our case, it places focus on the realization of the *VP*. Its general form is:

$$[\text{event } BECOME([I], [\text{event } GO_{+T}([I],$$

$$[\text{path } AT_{+T}([\text{VERB}([I], [J])])])])]$$

The  $GO_{+T}$  and the  $AT_{+T}$  characterize the evolution of the action to reach the resulting state via a kind of metaphorical path.  $T$  is the ontological domain of the resulting state. Finally, we leave the verb representation open, indicating only its two arguments  $I$  and  $J$ .

Typical example:

(1) *Les alpinistes ont atteint le sommet par ce chemin / The alpinists reached the top by this trail.*

[event  $CAUSE([event SELECT([alpinists],$   
 $[state BE_{+loc}([trail], AT_{+loc}([telic-of(trail, top)]))])]),$   
 $[event BECOME([alpinists], [event GO_{+loc}([alpinists],$   
 $[path AT_{+loc}([reach(alpinists, top)]))])])]$

The telic-of function produces here, for example, the predicate :  
 $go\text{-via}(trail, top)$   
since  $go\text{-via}$  is a predicate in the telic role of the noun *trail*.

#### 4.1.2 Grâce à. General form :

$\lambda I, \lambda K, \lambda J$

[event  $CAUSE([event UNDERGO([thing I], [state BE_{+T}([K],$   
 $[property telic-of(K, J)]))])],$   
 $[event BECOME([thing I], [state VERB([I], [J]))])].$

Typical example:

(2) *Le tourisme prospère grâce au Canal du Midi / Tourism thrives thanks to the Canal du Midi.*

The representation of this example is then:

[event  $CAUSE([event UNDERGO([thing tourism],$   
 $[state BE_{+char, +ident}([Canal du Midi],$   
 $[property attract(Canal du Midi, tourism)]))])],$   
 $[event BECOME([thing tourism], [state prosperous]))]]$

#### 4.1.3 Au moyen de. General form :

$\lambda I, \lambda J, \lambda K,$

[event  $CAUSE([event \vee state CONTROL([I], [state ACT([thing K],$   
 $[purpose telic-of(K, -) \text{ or } VERB \text{ if unexpected use}]))])],$   
 $[event CAUSE([I], [state INCH(VERB([I], [J]))])]$

INCH is a function of the LCS that produces the resulting state of an action. In this case, it is preferred to BECOME in order to strongly focus on the resulting state rather than on the process denoted by the verb, which is less prominent. The agentivity of the subject  $NP0$  is strongly marked in this representation by the primitive CONTROL.

Typical example:

(5) *Jean s'est brûlé au moyen d'huile chaude.* / *John burned himself by means of boiling oil.*

[*event CAUSE*([*event v state CONTROL*([*john*], [*state ACT*([*thing boiling oil*], [*purpose burn(boiling oil, -)*]))]), [*event CAUSE*([*john*], [*state burned*([*john*])]))])]

Here *brûler* (*burn*) in the second line is inferred from the compound *huile chaude* (*boiling oil*), not from the noun *huile* (*oil*) alone. We assume that *brûler* is inferred from the adjectives *warm* / *hot* / *boiling* applied e.g. to liquids like oil. This is a form of lexical inference, therefore *burn* is not in the telic role of the Qualia of *oil* (Pustejovski, *ibid.*).

(4) *Jean ouvre la porte au moyen d'un cric* (*case of an unexpected use of the instrument*) / *John opened the door with a jack.*

[*event CAUSE*([*event v state CONTROL*([*john*], [*state ACT*([*thing jack*], [*purpose open(jack, -)*]))]), [*event CAUSE*([*john*], [*state open(door)*]))])]

The unexpected use of the instrument representation occurs when the verb *VERB* is not prototypical. This situation can be characterized by the fact that neither the verb nor one of its synonyms or super- types (if any) is present in the telic role of the Qualia of the instrument. For example, in the telic role of *jack* there is nothing about its ability to open a door.

#### 4.1.4 Avec. General form :

$\lambda I, \lambda K,$

[*event CAUSE*([*event SELECT*([*I*], [*thing K*])]), [*event REACT*([*K*], [*state PROP(K)* if explicit or *TELIC – OF(K, -)*]))], [*event BECOME*([*I*], [*event VERB*([*I*])]))]

Typical example:

(6) *Jean s'est brûlé avec de l'huile chaude.* (*action performed unwillingly*) / *John burned himself with boiling oil.*

[*event CAUSE*([*event SELECT*([*john*], [*thing boiling oil*])]), [*event REACT*([*oil*], [*state burn(oil, -)*]))], [*event BECOME*([*john*], [*event burned(john)*]))]

Most of the representations given here make a heavy use of the telic-of function: this shows the quasi-systematic metonymic character of instrumental expressions. This is not surprising since the notion of instrument is largely related

to telicity. Some examples show that inference forms must be used instead of a reference to a telic predicate to get a correct representation of the utterance.

## 4.2 Towards a representation of the underspecified notion of instrumentality

Given the sub-event structure indicated in the general forms of the representations, we can now abstract over the representations to get the most generic notion of instrumentality. Formula:

$$(i) (e2 (e1 )) \rightarrow e3$$

is now expressed in LCS terms. Its abstract and under-specified form is:

$$\lambda I, \lambda K, \lambda J, [\text{event CAUSE}([\text{event } E2([I], [\text{eventstate } E1([K], [\text{property telic - of}(K, J) \text{ or } \text{VERB}]))])],$$

$$[\text{event } E3([I], [\text{state resulting - state(VERB)}]))]$$

with:

$$E2 = \text{UNDERGO / SELECT / CONTROL}$$

$$E1 = \text{BE / REACT / ACT}$$

$$E3 = \text{CAUSE / BECOME}$$

As can be noted, each preposition sense has its own selectional constraints and representation.

A major general difficulty not proper to our approach, and which extends to many prepositions and also to the semantics of adjectives, is the identification of the properties at stake. In this paper, we made the assumption that it can be extracted via a macro such as  $\text{PROP}(K)$ , or via the telic-of function of the Generative Lexicon. This obviously does not solve the problem. One of our aims will now be the development of reasoning procedures or pragmatic criteria that identify these properties. Some cases are quite straightforward while others are particularly difficult. We believe this is a good example of the interactions between semantics and pragmatics, and a contribution to this area.

Another direction, more basic, is to develop acquisition mechanisms that identify selectional restrictions from large sets of tagged  $PP_s$  denoting instrumentality. Although the results given above are based on a rather large number of examples, we believe that our approach is still fragile for very fine-grained semantic phenomena, where the overlap between the above cases are relatively important. Extensive corpus analysis may contribute, among others, to reach the adequate level of restrictions for each use.

## 5. Conclusion

In this paper, we proposed an analysis of the notion of instrumentality, going from the abstract notion to its lexicalizations via preposition senses. A symmetric movement has then been suggested, from the representation of ex-

amples in LCS with their application constraints to the under-specified representation of instrumentality, via abstract representations of preposition senses. This analysis shows the complexity of the notion and the necessity of using complex knowledge such as the one found in telic roles, among others. A considerable amount of work, systematization and development of examples (including metaphors and metonymies) remains to be done in this domain. This work has been developed on the basis of about 200 sentence samples. However, we believe that this work, through a concrete study of a complex notion, induces analysis, methods and semantic representation formalisms appropriate for developing a general framework for a proper preposition semantics.

This work is a first effort towards the definition of an accurate semantics for a number of preposition classes which have seldom being studied within a computational linguistics perspective. The next step is to study prepositions denoting instrumentality and manners. Similarly to prepositions denoting instrumentality, this study also involves related studies such as metonymic forms (treated here by calls to the telic role of the argument), compositionality (with the verb and the NP), the expression of selectional restrictions, and different forms of knowledge representation and inference, among which, as advocated here, the generative lexicon.

Besides an in-depth analysis of prepositions, our aim is to introduce such an approach in a number of applications where prepositions play or should play a major role. Let us first mention machine translation where it is often useful to go as deep as interlingua forms (Dorr et al 97) to get correct translations. Prepositions should in the future play a major role in knowledge extraction since the compound preposition + noun type is a clear and quite simple trigger of semantic information such as localization, manner, instrument, accompaniment or expression of an approximation (Cannesson et al. 01). Finally, let us mention the area of natural language generation where preposition choice, an aspect of lexicalisation, is a delicate task (see Benamara et al., this volume). It also interacts much with syntax, in particular with alternations as advocated above, and also with various forms of verbal incorporation.

We believe that such a detailed analysis of prepositions is useful to guarantee a certain level of quality and adequacy of computational linguistics applications which do not rely only e.g. on stochastic observations. Although prepositions have a certain semantic and syntactic autonomy, we also believe that their semantics must be investigated in close connection with the verb and the NP semantics.

## Acknowledgments

We thank the numerous native speakers that helped us to constitute a corpus of uses that allowed us to stabilize our analysis. We also thank two anonymous reviewers.

## Notes

1. This classification of prepositions according to their semantic “weight”, has been recently proposed by (Cadiot, 1997) and goes back to (Spang-Hanssen, 1963)

## References

Cannesson, E., Saint-Dizier, P. (2001), *A general framework for the representation of prepositions in French*, ACL01 WSD workshop, Philadelphia.

Berthonneau, M., Cadiot, P. (1991) (eds.), *Prépositions, représentations, référence*, Paris : Larousse.

Cadiot, P. (1997), *Les prépositions abstraites en français*, Paris : Armand Colin.

Olivier B, (1999), *Les constructions du verbe : le cas des groupes prépositionnels argumentaux*, Thèse de doctorat, Université Paris 7.

Dorr, B., Olsen, M.B., (1997), *Deriving Verbal and Compositional Verbal Aspect for NLP Applications*, proc. ACL'97, Madrid.

Jackendoff, R., (1990), *Semantic Structures*, MIT Press.

Levin, B., (1993), *Verb Semantic Classes: a Preliminary Investigation*, Chicago University Press.

Mari, A. (2003), *Polysémie et Décidabilité. Le cas de avec ou l'association par les canaux*, Paris: L'Harmattan.

Pinkal, M. (1985), *Logic and Lexicon*, Oxford : Oxford University Press.

Poesio, M. (1996) *Semantic Ambiguity and Perceived Ambiguity*, In K. van Deemter and S. Peters, (eds.), *Semantic Ambiguity and Underspecification*. Stanford: CSLI Lecture Notes, pp. 159-201.

Poncet-Montange, A. (1991), *A propos des noms d'instruments: relations entre forme et sens*, Lingvisticae Investigationes XV: 2, pp. 305-323.

Pustejovsky, J., (1995), *The Generative Lexicon*, MIT Press.

Saint-Dizier, P. (1998), *Alternations and Verb Semantic Classes for French*, in Predicative Forms for NL and LKB, P. Saint-Dizier (ed), Kluwer Academic.

Spang-Hanssen, E. (1963), *Les prépositions incolores du Français moderne*, Copenhague, GEC Gads Forlag.

Talmy, L. (1976), *Semantic Causative Types*, In M. Shibatani (ed.), *Syntax and Semantics 6: The Grammar of Causative Constructions*. New York: Academic Press, pp. 43-116.

Wierzbicka, A. (1992b), *Semantic Primitives and Semantic Fields*, in A. Lehrer and E.F. Kittay (eds.), *Frames, Fields and Contrasts*. Hillsdale: Lawrence Erlbaum Associates, pp. 208-227.

## Chapter 19

# PREPOSITIONS IN COOPERATIVE QUESTION-ANSWERING SYSTEMS: A PRELIMINARY ANALYSIS

Farah Benamara and Véronique Moriceau  
*IRIT, 118 Route de Narbonne 31062 Toulouse, France*  
benamara@irit.fr and moriceau@irit.fr

**Abstract** In this chapter, we show how the semantic properties of prepositions can be used within a logic based cooperative question-answering system. We focus on a subset of spatial and temporal French prepositions, outlining a naive Euclidean semantic interpretation relevant for our purpose. We then show how these representations are used in dedicated reasoning schemas such as relaxation or generalization based on a set of relations that classify each preposition according to its interpretation. Finally, we give some aspects of how prepositions are generated in natural language both during aggregation and lexicalisation. Results are integrated and evaluated within the WEBCOOP project, an intelligent, cooperative question-answering system.

**Keywords:** Spatial and Temporal Prepositions, Cooperative Question-Answering, Natural Language Generation.

### 1. Introduction

Prepositions have received little attention in the Computational Linguistics and Natural Language Processing (NLP) communities, however their use in applications, such as indexing, knowledge extraction or text summarization, is crucial but requires an in-depth syntax and semantics analysis to be of any use.

We investigate in this chapter a specific NLP task where prepositions play a predominant role, namely advanced question-answering (QA). In this application, semantic properties of prepositions have to be taken into account during question analysis, answer retrieval and, finally, during natural language

answer generation. As an illustration, let us consider the following natural language question :

*Give me trains to go from Paris to Toulouse at 11:30am,*

If there is no direct response to that query, an advanced question-answering system, where specific reasoning schemas are coupled with NLG techniques, can generate, for example, the following natural language responses :

*There is no Paris-Toulouse train at 11:30am. The nearest ones are:  
trains **around** 11:30am: at 11.20am and at 11.44am,  
trains **in late morning** : at 11.56am.*

In this example, the preposition *around* and the prepositional compound *in late morning* make explicit the retrieval mechanisms used (constraint relaxation), leading to appropriate answers.

Advanced reasoning for Question-Answering systems, as described in (Burger et al, 2002) (Maybury, 2004), rises new challenges since answers are not only extracted from written texts or structured databases but also constructed via several forms of reasoning in order to generate answer explanations and justifications. These systems require the integration of reasoning components operating over a variety of knowledge bases, encoding common sense knowledge as well as knowledge specific to a variety of domains by means, for example, of conceptual ontologies. Inference allows enhanced or extended QA services by providing intelligent and cooperative answers. There are many ways for a query to be answered intelligently, including :

- 1 providing answer explanations,
- 2 constructing intensional responses and answer summaries,
- 3 generalizing queries using neighborhood information (query relaxation),
- 4 comparing answers that have similar questions,
- 5 realizing information fusion when answers are inferred from multiple data sources, etc (Webber et al, 2002).

Our experimental framework is the WEBCOOP system (Benamara, 2004), a cooperative QA system, going from question processing to natural language answer generation using advanced reasoning and NLP procedures. To have a more accurate perception of how cooperativity is realized in man-man communication, we collected a corpus of French question answer pairs found in a number of web sites about tourism. An analysis of this corpus gave us interesting results on the way a cooperative system can be designed (Benamara et

al, 2004b). An analysis of this corpus also shows the prominent role played by prepositions to describe e.g. localization, instrument or purpose.

In this chapter, we show how the semantic properties of prepositions are used in WEBCOOP for providing intensional and relaxed answers which respectively correspond to the cases (2) and (3) cited above. We study in this framework a subset of temporal and spatial usages of prepositions. We elaborate a semantic representation for each preposition by means of a simplified version of the Lexical Conceptual Structure (LCS) and then we associate to each representation an interpretation function based on a naive Euclidean geometry, which is adequate for our purpose. These representations are then used in specific reasoning schemas in order to provide direct, relaxed or intensional responses. For this purpose, we define three relations namely *the inclusion*, *the opposite* and *the approximation* relations, that classify each preposition according to its interpretation. These relations depend on the ontological type of the object being localized and can be viewed as a set of axioms that allow, via cooperative reasoning procedures, for the retrieval of direct or additional responses that better answer a question. Finally, answers have to be generated in natural language. At the end of this chapter, we focus on general and on some specific aspects of preposition lexicalisation and aggregation of prepositional phrases.

## 2. Preposition use in Question Answering : the WEBCOOP system

Our general, experimental framework is the WEBCOOP system (Bena-mara, 2004), a logic based QA system that integrates knowledge representation and advanced reasoning procedures to generate intelligent and cooperative responses to natural language (NL) queries on the web in French. Cooperative responses are designed to respond to unanticipated questions and to resolve situations in which no direct answer is found in the data sources. In these cases, cooperative responses are useful to provide approximate answers, to make the comprehension of the response easier, to give explanations or to make suggestions.

Our approach requires the development of a knowledge extractor from web pages and the elaboration of a robust and accurate question parser. NL responses are produced from semantic forms constructed from reasoning processes. During these processes, the system has to decide, via cooperative rules, what is relevant and then to organize it in a way that allows for the realization of coherent and informative responses using the domain ontology and general knowledge. In WEBCOOP, responses provided to users are built in web style by integrating natural language generation (NLG) techniques with hypertexts links in order to produce dynamic responses (Reiter et al, 1997).

To evaluate the formalism and the cost of the procedures involved, the WECBOOP system is developed, in a first stage, on a relatively limited domain that includes a number of aspects of tourism (accommodation and transportation).

## 2.1 Prepositions in FAQ Corpora

To carry out our study, we considered three typical sources of cooperative discourses: Frequently Asked Questions (FAQ), Forums, and email question-answer pairs (EQAP), these latter obtained by sending ourselves emails to relevant services (e.g. for tourism: tourist offices, airlines, hotels). Our study was carried out on 370 cooperative question-answer pairs. The domains considered are basically large-public applications: tourism (45%), health (22%), sport, shopping and education (19%). In all these corpora, no user model is assumed, and there is no dialogue: QA pairs are isolated, with no context.

Prepositions in our corpora are generally used in questions to express a constraint and in the response to introduce a restriction, a direct or an additional answer. There are many ways to introduce a direct response.

One way is to use the same preposition as the one used in the question (*to whom? to someone*). When the type of the expected response is an entity (a name, a location, etc), the preposition used in the response belongs to the same semantic field as the one of the interrogative pronoun (*where? in*), like in the following QA pair:

*A qui s'adresser concernant l'utilisation d'informations provenant de (...)?*

*(Who should I talk to get information about...?)*

*Vous pouvez effectuer votre demande par mél à (...).*

*(You can send an e-mail to (...)).*

Another way is to use a preposition that belongs to the same semantic class as the interrogative pronoun. For example:

*Où se situent vos comptoirs et points de vente Aéris Express?*

*(Where are your Aeris Express desks?)*

*Ces derniers sont situés dans les aéroports suivants : (...)*

*(These are in the following airports: (...)).*

We have also noticed in our corpora that prepositions are used to introduce the context of the direct response. For example:

*Comment insérer une photo dans mon annonce?*

*(How can I insert a picture to my ?)*

*Pour insérer une photo, vous devez (...)*

*(To insert a picture, you have to (...)).*

Prepositional phrases are also used to introduce additional answers. In this case, prepositions describe the noun representing a direct response. For example, if the question: *le supermarché est-il loin de mon immeuble?* (*is the supermarket far from my building?*), has no answer, another preposition is used to describe the cooperative response, as in:

*Non, il se trouve à environ 50 mètres. Vous pouvez y aller à pied.*

(*No, It is at about 50 meters. You can go there on foot.*)

Another way to introduce a direct response is shown in this last example:

*Peut-on voyager avec son animal de compagnie ?*

(*Can we travel with our pet?*)

*Vous pouvez emmener votre animal de compagnie, si celui-ci a été vacciné contre la rage.*

(*You can travel with your pet only if it has been vaccinated against rage.*)

In this QA pair, the verb *emmener* used in the answer includes the notion of accompaniment, which is explicitly mentioned in the question, with an incorporation of the preposition *avec*.

In this chapter, we restrict ourselves to spatial and temporal usages of prepositions, which are the main prepositions found in our corpus. We show how these ones are used within WEBCOOP both during the reasoning steps and the natural language generation of cooperative responses.

### 3. Semantic Representation and Interpretation of Localization Prepositions in WEBCOOP

From a linguistic point of view, spatial and temporal prepositions have received an in-depth study for a number of languages (Verkuyl et al, 1992). The semantics of other types of prepositions describing manner, instrument, amount or accompaniment remains largely unexplored, see a general typology in (Cannesson et al, 2002). We describe in this section how we built our semantic representation.

#### 3.1 Semantic Representation

In (Cannesson et al, 2002), different senses of French prepositions are classified. The localization class includes the following facets: *source*, *destination*, *via/passage*, and *fixed position* which all get different semantic representations. These facets are considered as underspecified senses. From an ontological point of view, these facets can be applied to spatial and temporal arguments, as well as to metaphorical transpositions, e.g. into the psychological or the epistemic domains.

In this classification, each preposition sense is associated with a semantic representation by means of the Lexical Conceptual Structure (LCS) (Jackendoff, 1990). In (Cannesson et al, 2002), 35 LCS primitives are used to cover all the senses of the localization prepositions. These primitives correspond to concepts and are directly preposition names in the LCS meta-language. However, they are not necessarily used directly for the corresponding preposition. The LCS general forms are respectively for spatial and temporal prepositions:

$$\lambda X [_{\text{place/path}} \text{LCS Primitives}_{+\text{loc}} ([_{\text{place/thing}} X])] \text{ and,}$$

$$\lambda X [_{\text{place/time/path}} \text{LCS Primitives}_{+\text{temp}} ([_{\text{event/time}} X])]$$

where *place*, *thing*, *event*, *path* and *time* are conceptual categories and *+loc* and *+temp* are semantic fields of the LCS language.

In WEBCOOP, we use a simplified version of the LCS since conceptual categories and semantic fields are very restricted and can straightforwardly be inferred from predicate arguments. The simplified version is expressive enough and adequate to our needs in terms of knowledge representation and reasoning schemas. The LCS general form:

$$\lambda X [_{\text{place/path/time}} \text{LCS Primitives}_{+\text{loc}/+\text{temp}} ([_{\text{event/time/place/thing}} X])]$$

is then rewritten into the following predicate:

$$\text{LCS Primitives} (\text{SemField}, Y, X).$$

In the simplified form, *SemField* can be either *+loc* or *+temp* which corresponds respectively to a spatial or to a temporal preposition. The conceptual categories *place*, *thing*, *event*, *path* and *time* are omitted because, for example, they can be retrieved by access to the ontological type of the localization *X* or by access to a lexicon. The argument *X*, in the simplified notation, corresponds to a *geographical localization* for spatial prepositions and to an *event* or a *time unit* for temporal prepositions. The argument *Y* is the object being localized and it corresponds to one of the following concepts defined in our domain ontology: *tourist accommodation*, *tourist equipment*, *geographical localization*, *food place*, *transport infrastructure*, *means of transportation*, etc. These ontological types are also used to disambiguate a function w.r.t. its temporal or spatial usage. Here are some examples:

*L'hôtel est en face de l'aéroport d'Orly* (*the hotel is in front of Orly airport*) is represented by  $: \text{hotel}(X) \wedge \text{in front of}(\text{loc}, X, \text{orly}) \wedge \text{airport}(\text{orly})$

*Le train arrive entre 10 heures et 12 heures* (*the train arrives between 10am and 12am*) is represented by:

$$\text{train}(X) \wedge \text{arrivaltime}(X, Y) \wedge \text{between}(\text{temp}, Y, 10\text{am}, 12\text{am}).$$

In these examples, the predicates *infrontof* and *between* correspond to the semantic representations of the prepositions *en face de* and *entre*.

The following two tables summarize the semantic representations associated to each localization preposition found in WEBCOOP. A simplified version of the LCS is associated with each of them. A comprehensive LCS representation for some localization prepositions is given in the introduction of this book. In these tables, we give an approximate English translation for each French preposition (sense). Table 19.1 is dedicated to temporal prepositions and table 19.2 to spatial prepositions.

Temporal Prepositions	Simplified LCS Version
Après (after)	after(temp, Y, X)
A la fin de (at the end of)	end(temp, Y, X)
A (at)	at(temp, Y, X)
Au début de (at the beginning of)	beginning(temp, Y, X)
Avant (before)	before(temp, Y, X)
Au milieu de (middle of)	middle(temp, Y, X)
Vers, autour de, aux environs (around)	around(temp, Y, X)
Non loin de, près de (not far from, near)	near(temp, Y, X)
Entre (between)	between(temp, Y, X1, X2)

Table 19.1. Simplified LCS Representations for Some Temporal Prepositions

### 3.2 Interpretation of Localization Prepositions

There are many studies on how to interpret the meaning of spatial and temporal prepositions. Globally, various works related to psychology and linguistics such as those of (Herskovits, 1986) (Miller et al, 1976) (Talmy, 1983) and works in the field of naive physics (Hayes, 1976), show that property of space in natural language is different from the absolute spaces where entities are localized by means of coordinates. Thus, purely geometrical representation of the semantic of spatial prepositions is not appropriate.

However, in our case, we consider that staying at a geometrical level of representation involving e.g. geometry relations is sufficient for providing global reasoning schemas dedicated to cooperative responses (cf. section 4).

In WEBCOOP, we associate to each semantic representation an interpretation function based on a naive Euclidean geometry. We roughly consider that temporal prepositions are interpreted along a single dimension axis (time) and

Spatial Prepositions	Simplified LCS Representation
Après (after)	after(loc, Y, X)
A côté de (next to)	nextto(loc, Y, X)
A gauche de (at the left)	left(loc, Y, X)
A droite de (at the right)	right(loc, Y, X)
A la fin de (at the end)	end(loc, Y, X)
Au long de (along)	along(loc, Y, X)
Avant, en avant (front)	frontof(loc, Y, X)
Arrière, en arrière (back)	backof(loc, Y, X)
A l'est, à l'ouest, au nord, au sud (est, (west, north, south)	est(loc, Y, X), west(loc, Y, X), north(loc, Y, X), south(loc, Y, X)
En haut de (at the top of)	top(loc, Y, X)
En bas de (at the bottom of)	bottom(loc, Y, X)
Au centre, au milieu (at the center of, at the middle of)	middle(loc, Y, X)
loin de (far from)	farfrom(loc, Y, X)
A proximité, près de, proche de, auprès de, non loin de (near)	near(loc, Y, X)
A (at)	at(loc, Y, X)
Dans, en (in)	in(loc, Y, X)
Sur (on)	on(loc, Y, X)
Au début de (at the beginning of)	beginning(loc, Y, X)
Avant (before)	before(loc, Y, X)
Vers, autour de, aux environs, aux abords (around)	around(loc, Y, X)
Entre (between)	between(loc, Y, X1, X2)
En face de (in front of)	infrontof(loc, Y, X)

Table 19.2. Simplified LCS Representations for Some Spatial Prepositions

that spatial prepositions are interpreted in a space of one, two or three dimensions.

Let  $PRIM$  be the set of predicates associated to LCS primitives as described in the previous section (cf. table 19.1 and table 19.2). In WEBCOOP, we associate to each predicate  $prim \in PRIM$  an interpretation  $I_p$  defined by the pair  $(D_p, IF_p)$  where:

-  $D_p$  is the domain of interpretation that corresponds to a set of Cartesian coordinates in the Euclidean space. Each argument in  $prim$  (except the semantic field) is an element of  $D_p$ .

-  $IF_p$  is an interpretation function associated to each predicate  $prim$ . It is defined as follows :

$$IF_p : PRIM \rightarrow \{true, false\}$$

$$IF_p(prim) \rightarrow Z.$$

We consider that the objects being spatially or temporally localized are solid objects whose shape is roughly parallelepipedic, cylindrical or spherical. Let us now describe the interpretation functions associated to some preposition primitives. For the moment, those interpretations are relatively simple and are built according to our own intuition.

**3.2.1 Geometric Interpretation of Some Spatial Prepositions.** We study here the primitives: *farfrom*, *near*, *around* and *in*. Let  $Area_1$ ,  $Area_2$ ,  $Area_3$  and  $Area_4$  spaces of Cartesian coordinates respectively associated to each of these primitives, as shown in figure 19.1. Let  $Coord(Y)$ , a function that determines the coordinates of an object  $Y$ . Let also  $Interval_i(X)$ , the set of all coordinates of objects  $a$  that belongs to the space  $Area_i$  ( $i \in \{1, 2, 3, 4\}$ ) delimited by an object  $Y$ . Intuitively, we give the following interpretations:

$IF_p(in(loc, Z, X))$  is true if and only if  $Coord(Z) \in Interval_4(X)$ ,

$IF_p(around(loc, Z, X))$  is true if and only if  $Coord(Z) \in Interval_3(X)$ ,

$IF_p(near(loc, Z, X))$  is true if and only if  $Coord(Z) \in Interval_2(X)$ ,

$IF_p(far from(loc, Z, X))$  is true if and only if  $Coord(Z) \in Interval_1(X)$ .

In this formalization, we do not put forward any hypothesis about the geometric shape of the object  $Z$ .

### 3.2.2 Geometric Interpretation of Some Temporal Prepositions.

We study here the primitives: *at*, *around*, *near*, *after* and *before*. We represent temporal prepositions in a space of one dimension (time). Therefore,  $Coord(Z)$  is equivalent to  $Z$ . In this context and according to the figure 19.2,

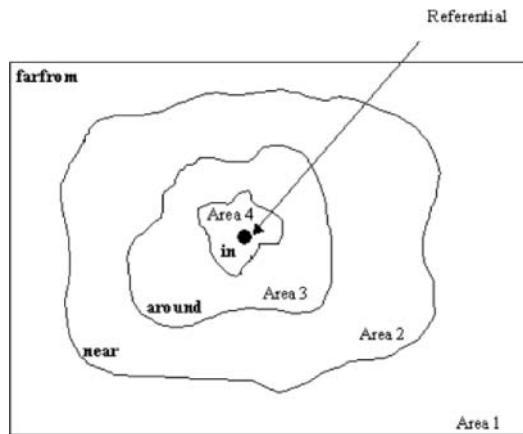


Figure 19.1. Geometric Interpretation of Some Spatial Prepositions

we give the following interpretation functions:

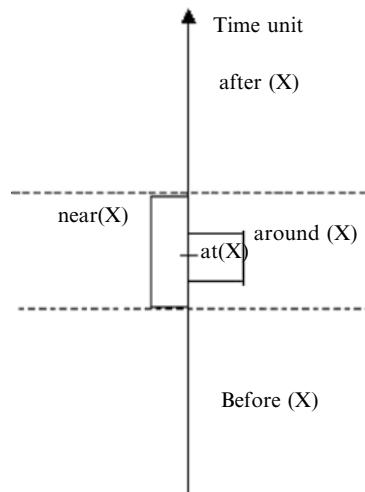


Figure 19.2. Geometric Representation of some temporal prepositions

$IF_p(at(temp, Z, X))$  is true if and only if  $Z = X$ ,

$IF_p(around(temp, Z, X))$  is true if and only if  $Z \in [X - n_1, X + n_1]$ ,

$IF_p(near(temp, Z, X))$  is true if and only if  $Z \in [X - n_2, X + n_2]$ ,

$IF_p(\text{after}(\text{temp}, Z, X))$  is true if and only if  
 $\exists Y > X, Z \in [X, Y] \text{ and } Z - X > n_3$   
 $IF_p(\text{before}(\text{temp}, Z, X))$  is true if and only if  
 $\exists Y < X, Z \in [Y, X] \text{ and } X - Z > n_3$

where  $n_1$ ,  $n_2$  and  $n_3$  are real numbers with  $n_3 > n_2 > n_1$ . For example, the expression *aux environs de 13h* can have the interpretation:

$IF_p(\text{around}(\text{temp}, X, 1\text{prn}))$  is true if  $X \in [12 : 50\text{am}, 1 : 10\text{pm}]$ .

The result of the interpretation function  $IF_p$  depends on the ontological type of both the considered object being localized and the localization. For example, the expression *the hospital is near the hotel* is different from the expression *the airport is near the hotel* because the proximity of the hospital is a walking distance w.r.t. to the argument hotel, whereas the proximity of the airport is rather evaluated in larger terms (e.g. kilometers). This is also the case in *the train arrives at 10am* which has a different interpretation from *the biker arrives at 10am*. In fact, the arrival hour of a public transport like a *train* implies a notion of punctuality that does not necessarily characterize a *biker*.

Spatial and temporal prepositions must be interpreted differently depending on the context. A common approach to this problem is the use of pre-specified interpretation values for each specific concept in the domain ontology. This approach requires experts to provide such information or to extend the interpretations to different domains (Minock et al, 1996). For the moment we give arbitrary interpretations to spatial and temporal prepositions used within WEBCOOP according to the semantics of the concepts we consider in our system. In the future, we plan to go further by using cognitive linguistics approaches (Dubois, 1989) coupled with more advanced geometrical considerations.

#### 4. Reasoning with Localization Prepositions

The above semantic representations are used to represent both (a) web pages and (b) NL questions. From the semantic representation of (a) and (b), an inference engine searches for appropriate responses and constructs the logical representation of a cooperative answer. A response is structured in two parts. The first part contains explanation elements in natural language. It is a first level of cooperativity that reports user possible misconceptions in relation with the domain knowledge. The second part is the most important and the most original one. It reflects the know-how of the cooperative system, going beyond the cooperative statements given in the first part. It is based on intensional description techniques and on intelligent relaxation procedures going beyond classical generalization methods used in artificial intelligence. This component

also includes additional dedicated cooperative rules that make a thorough use of the domain ontology and of general knowledge.

In this section, we show how semantic representations of prepositions are used in the know-how component. Before describing the reasoning schemas that we consider, we first present, in the next section, the prerequisite of our method. Then we present our reasoning strategy which is based on a set of relations that classify each preposition according to its interpretation.

## 4.1 Prerequisites

**4.1.1 Semantic Representation of Extracted Web Pages.** Each web text is represented under the form of a frame with attribute-value pairs where values can be atomic or fragments of LCS. The information extraction technique is based on a domain ontology where each major node is associated with a dedicated local grammar that recognizes, in the textual part of a web page, information relevant to the concept associated with that node.

The domain ontology is a concept ontology, where nodes are associated with properties. Concepts are connected to a lexicon to establish the link with lexical items and their syntactic properties. The lexicon reflects large fragments of the domain knowledge: what function or role an object plays, how an object or a complex event is organized, etc. It also contains selectional restrictions and semantic representations.

Here is the representation formalism produced by our extractor (Benamara et al, 2003). For each web page, we get the following logical representation:  $webtext(P_1(X_{11}, \dots, X_{1n}) \wedge \dots \wedge P_i(X_{i1}, \dots, X_{in}), http)$  where,

- $P_i$  are predicates that represent concepts in the domain ontology (unary predicates by convention), or their associated properties by means of n-ary predicates or semantic representations of prepositions (as described in table 19.1 and table 19.2)
- The arguments  $X_{ab}$ , (a=1 to i and b=1 to n) are variables or constants.
- $http$  is the web address associated to the indexed web page.

Our knowledge extraction system extracts fragments of information from web texts, as specified in the ontology. For example, consider the following text fragment :

*l'hôtel Sofitel à Paris possède une piscine et se situe en face de l'héliport de Paris*

*(the Sofitel hotel in Paris has a swimming pool and is situated in front of the héliport de Paris )*

if the extractor can fully parse it, it then produces the following logical representation:

$$\text{webtext}(\text{hotel(sofitel}) \wedge \text{equipment}(\text{sofitel}, Z) \wedge \text{swimmingpool}(Z) \wedge \text{in}(\text{loc}, \text{sofitel}, \text{paris}) \wedge \text{infrontof}(\text{loc}, \text{sofitel}, \text{heliport\_de\_paris}) \wedge \text{airport}(\text{heliport\_de\_paris}), \text{www.sofitelparis.fr})$$

In this representation, the predicate *hotel* and *swimming pool* are concepts in our domain ontology, *heliport de paris* is an instance of the concept *airport* and the predicate *equipment* is a property of the concept *hotel*. The predicates *in* and *infrontof* are semantic representations of the prepositions *à* and *en face de*.

**4.1.2 Semantic Representation of NL Queries.** NL questions are fully parsed and are represented using the same logical representation as web texts. We consider in this chapter questions that contain spatial and/or temporal constraints expressed by means of localization prepositions. For example, the NL question: *donnez-moi les gîtes en bord de mer en Corse* (*give me country cottages by the seaside in Corsica*) has the following logical representation:

$$Q = (\text{entity}, X : \text{countrycottage}, \text{countrycottage}(X) \wedge \text{region}(\text{corsica}) \wedge \text{sea}(Y) \wedge \text{in}(\text{loc}, X, \text{corsica}) \wedge \text{atborder of}(\text{loc}, X, Y))$$

where *entity* is the question conceptual category, *X* is the question focus, and the conjunction :

$$\text{country cottage}(X) \wedge \text{region}(\text{corsica}) \wedge \text{sea}(Y) \wedge \text{in}(\text{loc}, X, \text{corsica}) \wedge \text{atborder of}(\text{loc}, X, Y)$$

is the logical representation of the query body. In this representation, the predicates *in*(*loc*, *X*, *corsica*) and *atborder of*(*loc*, *X*, *Y*) are the localization constraints to be satisfied by the focus *X*.

Let us now show how the interpretation associated to each LCS primitives are used in WEBCOOP during the reasoning processes (section 4.2) and the generation of NL responses (section 5).

## 4.2 Reasoning with Prepositions: Our Approach

In this chapter, we consider the following reasoning schemas:

- **Case a)** Retrieval of direct answers when the question has a response in our knowledge base of extracted web pages. In this case, the system retrieves the list of answers that satisfy the question constraints. For example, if we have the following interpretations :

$IF_p(after(temp, X, 11am))$  true if  $X \in [11am, 24pm]$ ,  
 $IF_p(around(temp, X, 11 : 30 am))$  true if  $X \in [11 : 20am, 11 : 40am]$   
and  $IF_p(end(temp, X, [8am, 12am]))$  true if  $X \in [11am, 12am]$ ,

then the question: *give me trains to go from Toulouse to Paris after 11am?* can have the following reformulations: trains *around 11:30am* and trains *in late morning*.

- **Case b)** Retrieval of associated or neighborhood information when the question has an empty or a too small set of answers. In this case, an unsatisfied constraint expressed in the query by means of a localization preposition is relaxed into a preposition that better approximates the failed query. For example if the question: *give me hotels in front of Orly airport?* has no answer, the system can retrieve hotels *around* or *near* Orly airport. In fact, the interpretation associated to the expression *in front of Orly* is relaxed to take into account a larger area as generated by the interpretation of the expressions *around Orly* and *near Orly*.
- **Case c)** Generalization or summarization when the set of retrieved answers is large. In this case, the system tries to find a semantically more generic expression that better summarizes the list of extensional responses. For example, if the question: *list of palaces in Monaco with swimming facilities?* has the following possible responses:
  - *palaces near the sea,*
  - *palaces at less than 100 meters from the sea,*
  - *palaces in front of/by the sea,*

WEBCOOP can provide the summarized answer: *all Palaces in Monaco are near the Mediterranean sea* because the interpretation of the expression *near the sea* includes the interpretations of the expressions, *at less than 100 meters from the sea* and *in front of/by the sea*.

**4.2.1 WEBCOOP Reasoning Strategies.** To use localization prepositions in the WEBCOOP inference engine, we define different relations among both spatial and temporal prepositions that allow for the classification of each preposition wrt to another one according to its interpretation. These relations depend on the ontological type of the considered object and are not associated to any specific reasoning schema. They are defined according to a coarse-grained level which is actually adequate to our goals. These relations can be viewed as a set of axioms that allow, via cooperative reasoning procedures, for

the retrieval in the knowledge base of a direct (case a), a relaxed (case b), or a summarized response (case c) that better answers the question.

Let us consider two semantic representations  $Prep_a$  and  $Prep_b$  and their associated interpretations  $IF_p(Prep_a)$  and  $IF_p(Prep_b)$ . The general form of our classification relation R is  $Prep1 \ R \ Prep2$  where R can be one of the following relations:

- The *inclusion* relation:  $Prep_a \subset Prep_b$  means that  $Interval_a$  is included into  $Interval_b$  (cf. section 3.2.1) in terms of distance for spatial prepositions or temporal interval for temporal prepositions. For example, for an object  $X$  of ontological type village, we have:  
 $at(loc, Y, X) \subset around(loc, Y, X)$ .  
The relation  $\subset$  is reflexive, transitive but not symmetric. This relation is applied to both spatial and temporal prepositions.
- The *opposite* relation:  $Prep_a = Prep_b$  means that  $Prep_a$  has an opposite interpretation regards to  $Prep_a$ . For example  $right(loc, X, Y) = left(loc, X, Y)$  and  $infrontof(loc, X, Y) = backof(loc, X, Y)$ . The opposite relation is symmetric, not reflexive and not transitive. This relation is applied for both spatial and temporal prepositions.
- The *approximation* relation:  $Prep_a \rightarrow Prep_b$  means that  $Interval_b$  approximates  $Interval_a$  (cf. section 3.2.1). For example,  $after(temp, X, 11pm)$  and  $before(temp, X, 10pm)$  can approximate the expression  $between(temp, X, 10pm, 11pm)$ . The approximation relation is reflexive, not symmetric and not transitive. This relation is actually applied to temporal usages of prepositions.

Table 19.3 presents some axioms used in WEBCOOP to classify spatial and temporal prepositions via the inclusion, the opposite and the approximation relations defined above. For abbreviation, the notation :

$Prep_a \ R \ \{Prep_b, Prep_c\}$  where  $R \in \{\subseteq, \equiv, \rightarrow\}$  indicates that:

$Prep_a \ R \ Prep_b$  and  $Prep_a \ R \ Prep_c$ .

Let us now describe how these relations are used in the inference engine to provide direct and additional answers. Let Q be a query constrained by a spatial and/or a temporal preposition and formalized by :

$Q = Prep(SemField, Y, X) \wedge Q_2(X_{21},..X_{2k}) \wedge \dots \wedge Q_i(X_{i1},..X_{in})^1$   
where  $Q_2,..,Q_i$  are the subqueries that constraint the arguments  $X$  and  $Y$  and  $Prep$  is a semantic representation of a preposition. The inference engine processes the query against the knowledge base using one of the following reasoning schemas :

<b>Some inclusion axioms</b>
$\{at(temp, Z, X), in(temp, Z, X)\} \subset around(temp, Z, X)$
$around(temp, Z, X) \subset near(temp, Z, X)$
$\exists X < Y, z_0 \in [X, Y],$
$\{at(temp, Z, z_0), in(temp, Z, z_0)\} \subset between(temp, Z, X, Y),$
$\{around(temp, Z, z_0), near(temp, Z, z_0)\} \subset between(temp, Z, X, Y)$
$\{at(loc, Z, X), in(loc, Z, X), along(loc, Z, X)\} \subset nextto(loc, Z, X)$
$\{atborderof(loc, .., X), frontof(loc, .., X), on(loc, .., X)\} \subset nextto(loc, .., X)$
$nextto(loc, .., X) \subset around(loc, .., X)$
$around(loc, .., X) \subset near(loc, .., X)$
<b>Some opposite axioms</b>
$right(loc, Z, X) \Leftrightarrow left(loc, Z, X)$
$infrontof(loc, Z, X) \Leftrightarrow backof(loc, Z, X)$
$after(temp, Z, X) \Leftrightarrow before(temp, Z, X)$
$farfrom(loc, Z, X) \Leftrightarrow at(loc, Z, X)$
<b>Some approximation axioms</b>
$\exists x_0, near(temp, Z, X) \Leftrightarrow after(temp, Z, X)$ with $Z \in [X, X + x_0]$
$\exists x_0, near(temp, Z, X) \Leftrightarrow before(temp, Z, X)$ with $Z \in [X - x_0, X]$
$around(temp, Z, X) \Leftrightarrow near(temp, Z, X)$
$\{at(temp, Z, X), in(temp, Z, X)\} \Leftrightarrow around(temp, Z, X)$
$\exists X < Y,$
$beginning(temp, Z, [X, Y]) \Leftrightarrow \{before(temp, Z, X), middle(temp, Z, [X, Y])\}$
$\exists X < Y,$
$end(temp, Z, [X, Y]) \Leftrightarrow \{after(temp, Z, Y), middle(temp, Z, [X, Y])\}$
$\exists X < Y, \exists x_0,$
$between(temp, Z, X, Y) \Leftrightarrow before(temp, Z, X)$ with $Z \in [X - x_0, X]$
$\exists X < Y, \exists x_0,$
$between(temp, Z, X, Y) \Leftrightarrow after(temp, Z, Y)$ with $Z \in [Y, Y + x_0]$
$\exists z_0,$
$after(temp, Z, X) \Leftrightarrow before(temp, Z, X)$ with $Z \in [X - z_0, X],$
$\exists z_0,$
$before(temp, Z, X) \Leftrightarrow after(temp, Z, X)$ with $Z \in [X, X + z_0]$

Table 19.3. Some Axioms Used in WEBCOOP

■ **Case a)** the formula:

$Prep_1(SemField, Y, X) \wedge Q_2(X_{21},..X_{2k}) \wedge \dots \wedge Q_i(X_{i1}, \dots X_{in})$   
 is a **direct response** to the query Q if  $Prep_1 \subseteq Prep$ .

■ **Case b)** the formula:

$Prep_1(SemField, Y, X) \wedge Q_2(X_{21},..X_{2k}) \dots \wedge Q_i(X_{i1}, ..X_{in})$   
 is a **relaxed response** to the query Q if :  $Prep \subset Prep_1$  or  $Prep \rightleftharpoons Prep_1$  or  $Prep \rightarrowtail Prep_1$ .

■ **Case c)** if the query Q has a long list of extensional answers of the following form:

$Prep_1(SemField, Y, X) \wedge Q_2(X_{21},..X_{2k}) \wedge \dots \wedge Q_i(X_{i1},..X_{in})$   
 $Prep_2(SemField, Y, X) \wedge Q_2(X_{21},..X_{2k}) \wedge \dots \wedge Q_i(X_{i1}, ..X_{in})$   
 $Prep_n(SemField, Y, X) \wedge Q_2(X_{21},..X_{2k}) \wedge \dots \wedge Q_i(X_{i1},..X_{in})$

Then, if  $\exists Prep_k \in \{Prep_1,..Prep_j\} Prep_i \subset Prep_k$  for all  $i \in \{1, .., j\}$  ( $k \in \{1, .., j\}$ ),

then **summarized answer** is :

$Prep_k(SemField, Y, X) \wedge Q_2(X_{21}, ..X_{2k}) \wedge \dots \wedge Q_i(X_{i1}, ..X_{in})$

It is important to notice in this case, that if no generaliser is retrieved, then a summarized answer is not provided.

At this stage, the inference engine produces a set of logical forms for each cooperative response that better answers the query. These representations will be subject to several forms of reorganizations during the language generation phase in order to generate concise and fluid NL responses.

## 5. Generating Prepositions and PPs

In general, a generation task can be divided into two stages. The first step consists in content determination (deciding what to say) while the second step consists in the microplanning tasks i.e. deciding how to formulate the information in natural language. The contents of the response is produced by reasoning procedures, as shown above, in the form of a logical formula : this formula is composed of predicates representing concepts of the ontology and each concept can be lexicalised by one or several words, phrases, etc. Then, the microplanning tasks can, in turn, be divided into several subtasks such as (Reiter and Dale, 1997):

- **lexicalisation:** consists in choosing the appropriate lexical item, of the appropriate syntactic category for a given concept, represented by a fragment of a logical formula. This can be realized, for example, by using a lexicon preferably based on the domain ontology, where semantic representations of concepts are described,
- **aggregation:** consists in generating simple or complex noun phrases, prepositional phrases, propositions, coordination, etc., in order to produce syntactically correct sentences, possibly more concise and that avoids redundancies,
- **referring expression generation:** consists in deciding what expressions can be pronominalized and what expressions should be used to refer to entities.

The higher level tasks, such as discourse planning, are carried out on top of these more basic operations. In our case, discourse organization is directly handled by reasoning procedures. Consequently, we will not consider this level in this section.

In the following sections, we present how prepositions and PPs are generated in the microplanning tasks, especially during the lexicalisation and aggregation steps, and what the encountered problems are during these phases.

## 5.1 Prepositions and Lexicalisation

Natural language generation is a complex process involving a number of decisions at several levels. We investigate in this section the lexicalisation of prepositions and PPs. Lexicalisation and aggregation are two operations that work concurrently on the different predicates of a logical formula. Choices made on a certain point strongly influence the choices made on the others. Prepositions being a mediator between predicative terms and their arguments, their lexicalisation is quite delicate. The concepts that can be realized as prepositions are also quite diverse and include prelexical criteria (e.g. incorporation) as well as syntactic considerations (alternations, aspect or focus).

We first study general cases where preposition generation can be considered as quite straightforward. Then, we investigate in more depth cases where the lexicalisation of prepositions is more challenging.

**5.1.1 Direct Realizations.** In a number of cases (cf. cases a, b and c in section 4.2), preposition lexicalisation can be done in two ways: either by a direct “translation” of its semantic representation given in the lexicon or by a paraphrase of the prepositional phrase via several forms of transformations, in particular based on lexical semantics relations (e.g. synonymy) and asso-

ciated lexical inference rules. If we consider the logical representation of the cooperative response:

$(train(X) \wedge departurepoint(X, paris) \wedge arrivalpoint(X, toulouse) \wedge$   
 $departuretime(X, Y) \wedge around(temp, Y, 11:30am) \wedge city(paris))$

There are two possible lexicalisations for the predicate *around* which expresses here the notion of approximation:

- a direct “translation” :

*des trains Paris Toulouse vers / autour de /... 11H30*  
*(Paris-Toulouse trains around 11:30 am)*

- a paraphrase by a temporal adverb, via lexical inference rules:

*des trains Paris-Toulouse en fin de matinée*  
*(Paris-Toulouse trains in late morning)*

Both answers are semantically equivalent even if the second one is more difficult to implement because lexical inference rules have to be paired with the lexicon (an interval of time can be associated to a particular lexicalisation) so that we can interpret a concept in another way and formulate it using another lexicalisation.

**5.1.2 Towards more complex realizations.** In some particular cases, starting from the semantic representation of the response, we can choose to lexicalise or not a certain preposition. For example, if we consider the question:

*Quelles compagnies aériennes assurent la liaison Nice-Tripoli?*  
*(Which airline companies serve the line Nice-Tripoli?)*

and its semantic representation:

$(entity, Y : airline\_company, flight(X) \wedge departurepoint(X, nice) \wedge$   
 $arrivalpoint(X, tripoli) \wedge company(X, Y) \wedge airline\_company(Y) \wedge$   
 $city(nice) \wedge city(tripoli))$

A possible cooperative answer can be:

*not(webtext(flight(X) \wedge departurepoint(X, nice) \wedge*  
*arrivalpoint(X, tripoli), http))*

which corresponds to the NL response:

*Il n'y a pas de vol direct entre Nice et Tripoli.*  
*(There is no direct flight between Nice and Tripoli)*

In this case, there is no explicit prepositional concept in the semantic representation of the response but prepositions are inferred from concepts that

incorporate them (we could have said *there is no Nice-Tripoli direct flight*). Concepts such as *departurepoint* and *arrivalpoint* do incorporate respectively, and independently from each other, prepositions such as *from* and *to*. Considering together these two concepts can also generate the pattern *entre A et B* (between A and B) where A and B are respectively the departure and the arrival points.

A predicate can be lexicalised a priori non deterministically, i.e. in several ways. Some choices can then be filtered out at later stages, e.g. to agree with the selectional constraints imposed by another lexicalisation, or to implement a particular focus or connotation (Moriceau et al, 2005). Consider, for example, the contrast between the question and the answer (found in a corpus):

*Les animaux sont-ils acceptés dans le VVF?*

(*Are animals allowed in the VVF?*)

*Oui, les animaux sont acceptés au VVF.*

(*Yes, animals are allowed at the VVF*)

In the semantic formula of this question, **in**(...) is a predicate which can be lexicalised in French by *dans*, *en*, *à*, etc. In the example we have:

- a lexicalisation realized as *dans* in the question, which focuses on the VVF spatial facet of *in*,
- and in the response, considering some semantic properties of VVF, the lexicalisation of *in* by *à /au* (but not by *en* for lexical reasons), to better focus on the administrative facet of the VVF.

On the other hand, in the following example, only one lexicalisation for the preposition is acceptable:

*Donnez-moi un gîte en Midi-Pyrénées en bord de mer*

(*Give me a country cottage in Midi-Pyrénées on the seaside*)

*Il n'y pas de gîte en Midi-Pyrénées en bord de mer car...*

(*There is no country cottage in Midi-Pyrénées on the seaside because...*)

In the semantic formula of the question and the response, the concept represented by *in (loc, X, midi\_pyrénées)* can be lexicalised in French par *dans* or *en*. However, only the preposition *en* is possible (*en Midi-Pyrénées/ \*dans Midi-Pyrénées (but: dans la région Midi-Pyrénées)*). Lexical restrictions, here probably a metonymic use (name of the region replacing the region), impose the use of *en* instead of *dans*.

## 5.2 Prepositions and aggregation

Aggregation is usually described as a process which, besides forming PPs, propositions and other basic structures, improves text quality, avoids redund-

dancy (Reape and Mellish, 1999) by means of e.g. coordination, or making propositions clearer, or by using appropriate alternations when some arguments are vague, fuzzy or not explicit. Aggregation and lexicalisation are processes running in parallel and probably with some forms of concurrency: aggregation is strongly linked to the choices made during lexicalisation and vice-versa. For example, a property of a concept can be lexicalised as an adjective, a relative clause, or a prepositional phrase, etc. A problem, at this level, consists in organizing the arguments of a concept in a correct order taking into account a number of parameters which are illustrated below.

Let us assume that the starting point of the generation process is a simple logical form (which is a logical implementation of a simplified version of the LCS). The formula :

*chalet (X)  $\wedge$  person (Y)  $\wedge$  capacity (X, Y)  $\wedge$  N  $\leq$  10*

can be expressed in natural language (in French) by various forms, among which :

- 1 *La capacité d'un chalet est inférieure à 10 personnes* (litt.: *the capacity of a chalet is less than 10 persons*).
- 2 *La capacité maximum d'un chalet est de 10 personnes* (litt.: *the maximum capacity of a chalet is 10 persons*).
- 3 *Un chalet a une capacité inférieure à 10 personnes* (litt. *a chalet has a capacity less than 10 persons*).
- 4 *Un chalet (accueille / contient) (au maximum / moins de) 10 personnes* (litt. *a chalet welcomes/contains less than 10 persons*).

In these language realizations, note that focus is different for each proposition, that two terms of different syntactic categories - an adjective or a preposition - are used to express the operator  $\leq$  : *maximum* or *inférieure à*, and that the property *capacity* can be expressed by a noun or a verb (e.g. *accueillir*, *contenir*). These examples are 4 schematic natural language patterns of the above formula. They involve different lexicalisations and argument organizations i.e. different aggregation mechanisms, and result in slightly different global meanings.

Let us represent these utterances by a simplified grammar, that outlines the differences between surface forms. In our notation below, between [ ] is the focus, object denotes the concept 'chalet', relation is  $\leq$  and value is 10, between quotes are predefined NL terms, which are, for most of them, prepositions or verbs:

- 1 [property] 'de' object relation value.
- 2 [property + inverse relation] 'de' object 'est de' value.

3 [ object ] ‘a’ property relation value.

4 [ object ] lexicalisation as verb(property) inverse relation / relation value.

Note that the difference between 3 and 4 is mainly a different lexicalisation of the property, which then entails a different argument realization.  
ordered.

If we now consider alternations, such as the passive form, 4 above can be realized as: *un maximum de 10 personnes sont accueillies / \*contenues dans un chalet*.

The focus is then the relation in inverse form (maximum). Since the original subject is a metonymy (un chalet for the owner of a chalet), it must be realized as a locative PP introduced by the preposition *dans* (in) in the passive form (metonymy is no longer acceptable). The verb *contenir* cannot undergo this construction, therefore constraining the lexicalisation process to select the verb *accueillir*.

## 6. Conclusion

In this chapter, we have shown how prepositions are used in WEBCOOP, a logic based cooperative question answering system. The analysis of Frequently Asked Questions corpus shows the prominent role played by prepositions to describe e.g. localization, instrument or purpose. We study in this framework a subset of temporal and spatial usages of prepositions and we show how they are used in the inference engine of WEBCOOP for providing a direct, a relaxed or an intensional response. For this purpose, we elaborated a semantic representation for each preposition by means of a simplified version of the Lexical Conceptual Structure (LCS). We then associated to each representation an interpretation function based on a naive Euclidean geometry. These representations are then used in specific reasoning cases using three relations namely *the inclusion*, *the opposite* and *the approximation*, that classify each preposition according to its interpretation. In order to generate in natural language the cooperative answers resulting from reasoning procedures, we also presented some aspects of prepositions generation focusing on preposition lexicalisation and aggregation of prepositional phrases.

This work has several future directions among which we plan to:

- study the interpretation function of other spatial and temporal prepositions in order to extend our set of classification axioms.
- investigate the use of prepositions in other cooperative cases such as the retrieval of inferred responses or information fusion when answers are inferred from multiple data sources. In fact, this kind of cooperative responses require the use of different reasoning capabilities as well as the development of different strategies for natural language generation.

## Notes

1. We represent in this section only the logical formula of the query body

## References

Benamara F. (2004). *WEBCOOP, a Cooperative Question Answering System on the Web*, PHD Thesis, Paul Sabatier University, November 2004, Toulouse, France.

Benamara F., Saint Dizier P. (2004a). *Advanced Relaxation for Cooperative Question Answering*, Book Chapter in New Directions in Question Answering, Mark T. Maybury, editor, AAAI/MIT Press.

Benamara F., Saint Dizier P. (2004b). *Construction de réponses coopératives: du corpus à la modélisation informatique*, Revue Québécoise de Linguistique RQL, numéro spécial : TALN, corpus et Web.

Benamara F., Saint Dizier P. (2003). *Knowledge Extraction from the WEB: an Experiment and an Analysis of its Portability*, Vivek, volume 15, number 1, January.

Burger J., Cardie C., Chaudhi V. et al, (2002). *Issues, Tasks and Program Structures to Roadmap Research in Question Answering*, Technical report, NIST.

Cannesson E., Saint-Dizier P. (2002). *Defining and Representing Preposition Senses: a Preliminary Analysis*, ACL WSD workshop, Philadelphia.

Dubois D. (1989). *Psychologie du langage, psycholinguistique, psychologie cognitive. Contribution de la psychologie aux sciences du langage*, Histoire, Epistémologie, Langage, numéro spécial sciences du langage et recherches cognitives. Pages 85-104.

Jackendoff R. (1990). *Semantic Structures*, MIT Press Cambridge M.A

Hayes P. (1985). *The second naive physics manifesto*, in: Hobbs – Moore (eds.), pages: 1-36.

Herskovits A. (1986). *A Language for Spatial Cognition*, Cambridge University Press.

Maybury M. (2004). *New Directions in Question Answering*, Mark T. Maybury, editor, AAAI/MIT Press.

Miller G., Johnson-Laird Ph. (1976). *Language and perception*, Cambridge, MA: Belknap Press of Harvard University Press.

Minock M., Chu W., Yang H., Chiang K., Chow G. and Larson C. (1996). *CoBase: A Scalable and Extensible Cooperative Information System*, Journal of Intelligent Information Systems, volume 6, number 2/3, pages: 223-259.

Moriceau V., Saint-Dizier P. (2005). *A Constraint-Based Model for Lexical and Syntactic Choice in Natural Language Generation*, Lecture Notes in Artificial Intelligence (LNAI), volume 3438.

Reiter R., Dale R. (1997). *Building Applied Natural Language Generation Systems*, Journal of Natural Language Engineering, volume 3, number 1, pages:57-87.

Reape M., Mellish C. (1999). *Just What is Aggregation Anyway?*, Proceedings of the 7th European Workshop on Natural Language Generation, Toulouse, France.

Talmy L. (1983). *How Language Structures Space*, in Pick – Acredolo (eds.), pages: 225-282.

Webber B., Gardent C., Bos J. (2002). *Position Statement : Inference in Question Answering*, LREC.

# Index

$\lambda$ -calculus, 296  
Agglutinative language, 69  
Alternation, 8  
Antonymy, 171  
Argument composition, 187  
Attachment tendency, 83–84, 87, 93  
  
Basque, 71  
Bisemic, 216  
Blocking  
  gap, 28  
Blocking gap, 39  
Bound word, 181, 190  
British National Corpus, 164  
  
Case, 105  
Category mistake, 236  
Causal relation, 267, 269, 271, 274, 278, 285  
Causality, 284  
Cause, 272, 274  
Channel, 276  
Circumposition, 92–93  
Classification, 14, 275  
Clause boundary, 83, 87  
Cofinal, 102  
Cognitive notion, 293  
Coinitial, 102  
Collocation, 181–182  
Collocational prepositional phrase, 182  
COMLEX, 164  
Complex lexical category, 183  
Complex preposition, 181  
Compositional semantics, 181  
Compositionality, 169, 234  
Conceptual categories, 295  
Conflation, 58  
Constraints, 275  
Contact predicates, 133  
Context dependent interpretation, 249  
Contracted preposition, 83–84, 87–88, 93  
Coordinated behavior, 274  
Coordination, 34, 274, 280  
Corpora, 121  
Countability, 163  
Counterfactuality, 272, 279, 284  
  
Data sparseness, 122  
Decomposable, 190  
Dependence, 263, 272, 284  
  
Determinerless PP, 163  
Dictionaries, 118  
Directional, 102  
Disambiguation, 84, 86–88, 234  
Distributional idiosyncrasy, 181  
Distributional similarity, 197  
  
English Resource Grammar, 173  
Event, 291  
Eventualities, 274  
External selection, 181, 190–191, 194  
  
False colouring, 217  
Flexible Montague Grammar, 186  
Flexible semantics, 186  
Formal ontology, 231  
Frame of reference, 214  
Frame of reference ambiguity, 214  
  
Gap, 38, 40  
Generative Lexicon, 302, 253  
Generative ontology, 236  
Genitive, 182–183, 192  
German preposition, 83–85, 93  
  
Head-complement rule, 174  
Head-driven Phrase Structure Grammar, 131, 173, 181–183, 186, 188, 190, 193–194, 247  
  
ID Principle, 192  
ID Schema, 192  
Idiosyncratic, 192, 194  
Impingement verbs, 135  
Incorporation, 58  
Indirect prepositional arguments, 131  
Infomorphism, 276  
Institutional nouns, 166  
Instrument, 263, 265–266, 273, 279–282  
Instrumentality, 293  
Intransitive preposition, 198  
  
Japanese Phrase Structure Grammar, 247  
Japanese postposition, 246  
  
Kullback Leibler Divergence, 167  
  
Landmark, 212  
Latent semantic analysis, 197  
Levin's verb classes, 124  
Lexical Conceptual Structure, 16, 72

- Lexical Semantics Templates, 295
- Lexicalisation, 173, 290
- Lexicalization, 58
- Lexicalized Flexible Ty2, 186–187
- Lexicon Principle, 192
- Linguistic Interaction with Virtual Environments, 212
- Liquid deletion, 31–32
  - vowel fusion, 31
- Locative expression, 212
- Manner, 263, 265, 273, 279–281, 283
- Markedness, 165
- Maximized Discourse Coherence (MDC) principle, 250
- Minimal Recursion Semantics (MRS), 131
- Mode, 102
- Modifiability, 165
- Motion verb, 109
- Multilingual context, 131
- Multiword expression, 163
- Neighbourhood, 102
- Non-decomposable, 194
- Notion of instrument, 289
- Notions, 264
- Object occlusion, 216
- Objects, 274
- Ontological semantics, 237
- Ontological type, 299
- Ontological well-formedness, 236
- Particle, 198
- Path, 58
- Phrasal lexical entry, 181, 193–194
- Phrasal lexical item, 191
- Phrasal verbs, 4
- Possessive construction, 257
- Postposition, 90, 92–93, 1, 69
- Posture verb, 109
- Potential field model, 215
- PP-attachment, 7
- Preposition semantics, 197
- Preposition valence, 197
- Prepositional phrase, 83
- Prepositional verbs, 4
- Primary preposition, 85, 87, 89–90
- Primitives, 17, 295
- Productivity, 165
- Proform, 183
- Pronominal adverb, 83–85
- Pronominal adverb, 89
- Pronominal adverb, 89–93
- Properties, 275, 266
- Raising, 181–184, 186
- Ray casting, 219
- Reciprocal pronoun, 83–85, 90, 93
- Relational compositionality, 234
- Removal predicates, 134
- Roget's thesaurus, 197
- Scheme, 279–280
- Search axis, 213
- Secondary preposition, 85
- Segmented Discourse Representation Theory, 249
- Selection, 163
- Semantic category, 291
- Semantic roles, 238
- Shifting operation, 186–189
- Sign, 104
- Situation, 271, 274, 277–278
- Spatial expression, 101
- Spatial nouns, 64
- Spatial template, 212
- Spatial template's origin, 213
- Subcategorization frame, 5
- Synonymy, 172
- Thematic roles, 6
- Trajector, 212
- Transitive preposition, 198
- Treebank, 84, 90, 93
- Type, 275
- Unboundedness, 166
- Underspecification, 16
- Underspecified form, 280, 291
- Unrealized argument, 184, 188
- Verb particle construction, 198, 115
- Verbal complex, 186