



Universiteit Utrecht

FACULTY OF THE HUMANITIES
DEPARTMENT OF PHILOSOPHY
COGNITIVE ARTIFICIAL INTELLIGENCE

MASTER'S THESIS

Bare PPs from a Multilingual Perspective

Author:

Martijn van der Klis

3020371

30 ECTS

Reviewers:

Prof. Dr. Henriette de Swart

Dr. Joost Zwarts

Prof. Dr. Albert Visser

July 29, 2010

Acknowledgements

Writing a thesis is by no means an easy feat. In my case, I often found myself surrounded with on the one hand an enormous amount of data and on the other hand a nice set of conclusions, waiting to be demolished by yet another data point. Nevertheless, I hope I have managed to find my way through at least some of the conundrums surrounding a both monolingual and multilingual analysis of bare prepositional phrases. Yet, I could not have done so with a little help of my friends, whom I in this (far too short) section would like to thank.

First of all, I owe a big 'thank you' to Henriette de Swart. As first supervisor, Henriette was able to create a very stimulating environment, always seeing the bigger picture out of the millions of tables and data points I provided at each meeting. Her inviting of both Joost Zwarts and Bert Le Bruyn to the meetings was a brilliant idea, as both Joost and Bert again and again provided me with interesting niches for me to work out. Joost's knowledge of the tricks of the trade is almost unparalleled, and as a second supervisor he provided me with some very important hints. Meeting with Bert also was a great pleasure, as his enthusiasm for (corpus) linguistics lightened me up every time, both literally and metaphorically.

Bert's enthusiasm made another 'victim' in Marten Postma, a student in French, with whom I was able to extend the analysis of 'per' and 'without' to French during two or three very nice and interesting sessions. Do read his Bachelor's Thesis for some very interesting results, of which I have included only a few in my thesis. Special thanks also to Albert Visser, who was so kind to take up the job as third reviewer and in passing gave me some important pieces of advice for future work on the semantics of bare PPs.

And then there of course are people outside the university that gave me some brilliant ideas or, sometimes even more important, kept me going with friendly words. Doing research into linguistics is not really understood by the outside world (remember Helen de Hoop's confrontation with Ronald Plasterk on the use and grammaticality of the Dutch 'hun' in subject position?), and I thus can not thank Daphne, Marja, Astrid, Kevin, Peter and Bernd enough for their time with me. And of course, I must thank my very significant other, Anouschka, who was actually quite like the bare PPs I have analyzed throughout my thesis: every time I see her, I always see something new to wonder about. I hope some of this excitement and wondering (about bare PPs, I mean) is to be found for the reader in this thesis.

Contents

1	Introduction	7
1.1	Bare PPs: what, when and why?	7
1.1.1	Between Idiomaticity and Productivity	8
1.2	Research Questions	9
1.3	Research Method	10
1.3.1	Corpus Linguistics: An Introduction	10
1.3.2	Finding Bare PPs	11
1.4	Relation to Artificial Intelligence	11
1.5	Review of the Literature	12
1.5.1	Distribution	12
1.5.2	Semantics	14
1.6	Outline of this Thesis	20
I	The Monolingual Perspective	21
2	Evidence from English	22
2.1	Introduction	22
2.1.1	Why Again English?	22
2.1.2	Method	22
2.1.3	Drawbacks to the Method	23
2.1.4	Two kinds of bare PPs	23
2.2	P-based bare PPs	24
2.2.1	<i>Per</i> : a UNITER	27
2.2.2	<i>Under</i> under study	27
2.2.3	<i>In</i> and <i>By</i> : a pair of antonyms?	29
2.2.4	Conclusion	30
2.3	\exists -based bare PPs	31
2.4	N-based bare PPs	32
2.4.1	Locations	32
2.4.2	Media	33
2.4.3	Parts of Body	35
2.4.4	Times	36
2.4.5	Conclusion	36
2.5	Conclusion	36
3	The Dutch Data	39
3.1	Introduction	39
3.1.1	Method	39
3.1.2	Drawbacks to the Method	39
3.2	P-based bare PPs	40
3.2.1	<i>Per</i> as Polysemous Preposition	42

3.2.2	<i>Te</i> : Bare	43
3.2.3	Activity Readings with <i>Op</i>	44
3.3	\exists -based bare PPs	44
3.4	N-based bare PPs	46
3.4.1	Locations	46
3.5	Conclusion	47
 II The Multilingual Perspective		49
 4 Statistical Analysis of the Multilingual Corpus		50
4.1	Introduction	50
4.2	A Determined Hypothesis	50
 5 Multilingual Analysis per Preposition		53
5.1	Introduction	53
5.2	Peculiarities of <i>per</i>	53
5.2.1	Method	53
5.2.2	Results	54
5.2.3	A short note on 'per + MEANS OF TRANSPORTATION'	59
5.2.4	Conclusion	61
5.2.5	Suggestions for further research	61
5.3	<i>Without</i> Revisited	62
5.3.1	Method	62
5.3.2	Results	63
5.3.3	Intermediate Conclusion	67
5.3.4	A New Method	67
5.3.5	Results	67
5.4	Conclusion	69
 6 Multilingual Corpora: Advantages and Technical Aspects		70
6.1	Introduction	70
6.2	Advantages to Comparable Corpora	70
6.3	Alignment	71
6.3.1	Sentence Alignment	71
6.3.2	Word Alignment	73
6.3.3	Conclusion	73
 7 Concluding Remarks		74
7.1	Summary of the Results	74
7.2	Suggestions for Further Research	75

Chapter 1

Introduction

1.1 Bare PPs: what, when and why?

This thesis will be all about **bare prepositional phrases**. Bare PPs are prepositions followed by *singular count nouns* lacking an overt determiner. A (made-up) example is (1.1), in which the bare PPs are printed in bold:

(1.1) I was told **by e-mail** that Pat was transferred **to prison, without discussion**. His case is now **under study**, and will be decided **in court**, probably **at noon**.

Due to the missing determiner, the semantics of the bare PP is often quite different from that of the full PP. An enlightening minimal pair, from [Stvan, 1998], is the following:

(1.2) Pat is in prison.

(1.3) Pat is in **the** prison.

While in (1.2), Pat is incarcerated, in (1.3) it might very well be that she is only in the prison on visit. Likewise we find that there is a distinction between "I am going to school" and "I am going to the school". The former denotes going to a school to perform a specific activity (learning stuff, most likely), while the latter can be uttered by a mother who is going to pick up her children. The difference in the semantics of both expressions can also be assessed when one tries to refer back to the noun in question. We see that in (1.5), a discourse referent is set up, but in (1.4), this seems not to be the case. (examples again from [Stvan, 1998]):

(1.4) Pat is in prison. #It is a 3-story concrete building.

(1.5) Pat is in **the** prison. It is a 3-story concrete building.

As a third difference between the two forms, we can look at their ability to occur modified. With (1.6), we see that the types of modification are restricted. With modifiers like 'big' and 'red', the bare PP as a whole is rendered ungrammatical. Only within a specific range of prepositions the bare PP is sensible and also keeps its enriched meaning. In (1.7) on the other hand, we see that the non-bare form allows free modification of the noun (the examples here are my own, checked for consistency with a Google search).

(1.6) Pat is in *big / *red / federal / state prison.

(1.7) Pat is in **the** big / red / federal / state prison.

From the observations above, we can conclude that there are two forms (a bare form in (1.2) and a non-bare form in (1.3)) and that that these two forms have different properties with regard

to semantics, the setting up of a discourse referent and the possibility of modification. However, in the domain of bare and non-bare (or better, *full*) PPs, things are not always that clear-cut.

To make this latter point, let us look at the next examples (my own, checked for consistency with a Google search). While the non-bare form in (1.9) is perfectly acceptable, the bare alternative in (1.8) is clearly ungrammatical. And in (1.10) and (1.11), we find that this time, it is the bare form that is the only viable alternative. We thus do not always find a competition in form and meaning.

(1.8) * Mary is at supermarket.

(1.9) Mary is at **the** supermarket.

(1.10) The thesis is under discussion.

(1.11) * The thesis is under **the** discussion.

Even though in most of the previous examples we found the bare PPs always occurring in object position (and thus being governed by a verb), this is in no means obligatory. Bare PPs can appear all over a sentence, as our very first example (1.1) showed us.

But why would one research bare PPs? Well, first of all, bare PPs are used widely in language. While people seem to deal quite effortlessly with these, for a computer, is it by no means obvious when a noun can appear bare after a preposition. If we would like to have a computer correctly predict whether a noun can appear bare, we will have to extract these rules from the data.¹

Next to this practical purpose, there is also some theoretical purpose of looking at bare PPs. The rules we extract for bare PPs, might be of help when we look at bare nouns in other positions of the sentence. Researching bare PPs can thus be of help completing the whole picture of bare nominals.

As a last argument for researching bare PPs, I argue that bare PPs form a nicely closed domain in which we can compare various languages. While for example scopal properties are difficult to compare between languages, bare PPs (1) are not *that* hard to find and (2) can be analyzed outside of a sentence's context (that is, most of the time, the context does not influence the grammaticality of the bare PP).

Bare PPs thus form an extremely interesting domain of research. While examples like (1.2) and (1.3) have been thoroughly studied by both [Stvan, 1998] and [de Swart and Zwarts, 2009a], there are many classes of bare PPs which have not yet been documented in the literature. I believe (1.10) is one of these classes, and in this thesis we will systematically explore whether there are more of these yet unexplored forms. In a way, this thesis is consequently sort of an adventure. But, as any adventure, this adventure is not without danger. Let us therefore first see which obstacles we will have to avoid.

1.1.1 Between Idiomaticity and Productivity

As said in the previous section, we are interested in bare PPs: prepositions directly followed by a bare singular count noun. However, there are actually quite a lot of forms which do follow this pattern, but still are not of any interest for our present purposes: they form *idioms*. A by no means conclusive list is provided below in (1.12):

(1.12) in fact, of course, for example, with respect (to), by means (of), on top (of), at stake.

What we are looking for, on the other hand, are productive forms of bare PPs. An example of such a productive case can be found in (1.13) below (from the second chapter, section on 'by'). We see here that we can insert every noun of the category MEDIUM into the slot created by 'by', rendering a reading in which something is communicated by the given noun.

¹Or we could try and let the computer find these rules for itself by using, for example, a neural network to automatically match input to output. However, such an approach seems quite unfruitful for something so structured as language, as for example [Fodor and Pylyshyn, 1988] have argued again and again.

(1.13) by MEDIUM: by (tele)phone, by letter, by radio, by (air)mail, by book, by postcard.

But how do we *prove* a certain class of bare PPs is productive? In my opinion, one has several ways to perform such an analysis, given that one is in the possession of either a corpus or an Internet search engine (Google, for example). A few of them are listed below. In this thesis, I have most of the time employed the Semantic Substitution test, not because of its validity, but mainly for its ease of use.

- **Semantic Substitution test:** when one finds an example $[P N_x]$ in a corpus, substitute N_x by a semantic equivalent N_y .² If one is able to find that $[P N_y]$ is grammatical (for example by attesting whether $[P N_y]$ occurs in another corpus or in the big corpus that is the Internet), one can say that the class is productive.
- **Corpus Split test**³: Split a corpus in three separate parts⁴, with sizes in a ratio of 1:2:3. Each part ideally consists of random lines from the corpus. Search for each part of the corpus for $[P N]$, with N the set of semantic equivalents $\{x, y, z, \dots\}$. Note the amount of types found in each part of the corpus. When one observes that the ratio of number of types found also is of ratio 1:2:3, one can say that the class is productive.

Given the previous examples, with bare PPs, we are thus in a continuous fight between idiomatycity and productivity. Of course, we are interested in mapping the second category, but first of all, how are we going to find the right data? In the methodological section of this introduction, I will comment on that matter. For now, let us continue with formulating the research questions of this thesis.

1.2 Research Questions

In this section, I will give an overview of the research questions under study in this thesis. First of all, in the introduction of this chapter, it became clear that we have not yet the slightest idea of when one can leave out the determiner to form a grammatical and non-idiomatic bare PP. We thus would like to know just what are the productive cases of bare PPs. And furthermore, we would like to know whether it is the preposition or the noun which licenses the bare occurrence. This leads to two main research questions for a monolingual analysis:

(1.14) Which prepositions license (categories of) bare nominals?

(1.15) Which (categories of) bare nominals can be found across a lot of prepositions?

In this thesis, I will first of all answer the two above questions for English. In the chapter after that, I will provide a summary of the results that [Le Bruyn et al., 2009] found for Dutch and add some of my own research. English and Dutch are of course by no means representative of the whole language spectrum. However, due to their almost similar usage of both the definite and indefinite article (see for example [de Swart and Zwarts, 2009b]), we are able to effectively compare the two languages when they decide to leave an article out. It would be much harder to compare English with a language like Romanian, which, as for example [Mardale, 2006] noted, can drop definite articles inside a PP with no change in the semantics of the expression.

Thus, English and Dutch will be the main languages under study. But how are we going to compare these two relatives? One could opt for simply comparing the results of the first two chapters or using native speakers to give a judgment on direct translations of the forms found in one of the two languages. While this latter approach has often been the way to go, in this thesis, I

²Semantic equivalents are here not necessarily nouns with the similar meaning, but may also be nouns that share a lot of properties with N_x .

³This test is a spin-off of the more valid test performed by [Dömges et al., 2007]. However, the currently proposed test is much more easy to perform, and in my opinion, the results will not deviate *that* much from the original.

⁴With Google, one might opt for the 'site:' tag to split the (enormous) corpus into several pieces.

would like to do something quite new⁵: I will use a corpus consisting of both Dutch and English versions of proceedings of the European Parliament, and confer these texts in their use of bare PPs. In this way, we are able to gather both syntactic and semantic evidence of commonalities and differences between English and Dutch. This leads to two research questions (and each can be asked for each of the two categories of bare PPs mentioned above):

(1.16) In a cross-linguistic analysis, do we find differences between languages with regard to bare PPs?

(1.17) How can we account for these differences (especially when they have a similar usage of articles)?

With answering the first two research questions, we are able to make a big advance in the research on bare PPs from a monolingual perspective. Furthermore, while giving an answer to the second two questions, we will pave the way for multilingual syntactical and semantical analysis concerning bare PPs, an area of research which (to my knowledge) has been neglected for some time now (with the exception of a small chapter in [Stvan, 1998]).

Having posed the research questions, it is now high time to explain *how* these are going to be answered. While I have already hinted at some of the research methods in before, I will dive deeper into this in the next section.

1.3 Research Method

In this section, I will first introduce the main research method in this thesis and its merits and disadvantages. After that, I will review in more detail how the data in this thesis have been gathered.

1.3.1 Corpus Linguistics: An Introduction

The research in this thesis is mostly **corpus linguistics**. In both past and present day, corpus linguistics has been both hailed and detested as an approach to linguistics. One of the main downsides of using the corpus as entity under study is the always lingering possibility of missing data due to underrepresentation in the corpus. Even when one employs a corpus of more than 100 million words, this possibility remains, due to the productivity and variability of human language.

On the other hand, using one's own imagination and judgments to come up with examples is often not enough to drive the interesting roads without losing your way. In my opinion, a corpus can often be a handy guide on your journey through both the syntactic and semantic landscape. This can be either by showing you ways you have never traveled before, or by giving you an idea of whether a certain way can be taken (checking the grammaticality of a sentence or construction).

In an important sense, this approach differs from earlier corpus linguistics. In these studies, the goal of the research was to describe the corpus: the corpus formed the entity under study. But in my opinion, corpora should not be perceived as being truthful in any way, no, I think one should look upon them more pragmatically as a tool to help a linguist find new ways to perceive the linguistic data. The data does not necessarily have to come from the corpus under study, but might also be found in other, bigger corpora (i.e. search engines) or one's own imagination (of course, then always to be checked by native speakers or again, search engines).

At the moment, corpus linguistics is regaining interest due to the upcoming multilingual corpora. In these corpora, we often find aligned text of two or more languages. Most of the time, these corpora have been applied in machine translation. But as for example [Johansson, 2007]

⁵Although, of course, I am certainly not the first using a multilingual corpus to find syntactic and semantic evidence: [Johansson, 2007] devoted a very inspiring book to methods and case studies in this kind of contrastive analysis.

shows, multilingual corpora are also ideal for both syntactic and semantic research across languages. In this thesis, we will make some first attempts at finding interesting themes within the domain of bare PPs using a multilingual corpus.

1.3.2 Finding Bare PPs

In this thesis, as stated before, I will use both monolingual and multilingual corpora to acquire results. But how to actually find bare PPs inside a big corpus? First of all, an important prerequisite for the corpus is that it is *tagged for parts of speech* (often abbreviated as POS-tagged). When this is the case, we are able to effectively search for prepositions followed directly by a singular noun by using regular expressions. A second preliminary would be that the corpus should be sufficiently big. Bare PPs are not rare (see the first section of this introduction), but to assure they appear frequently enough to provide workable statistics, one should have a corpus of at least a half million words.

For a monolingual analysis of English, we find that the Brown Corpus (as composed by [Kucera and Francis, 1967]) is a suitable candidate. The corpus consists of various written articles, tales and poems, has 1.014.312 words in total and is indeed tagged for parts of speech. For Dutch, the Eindhoven Corpus ([van Grootheest, 1989]) is used to find the bare PPs. This corpus consists of both written texts and transcribed conversations, has around 768.000 words in total and is also tagged for part of speech. In the multilingual part of this thesis, we will employ the EuroParl corpus ([Koehn, 2005]). While its size is more than sufficient (35 million words), it (unfortunately) lacks POS-tags.

While POS-tags indeed help to smoothen the search, the method of just searching for any preposition followed by any singular noun is by no means fool-proof, as will be shown in the more in-detail methodology section of the second chapter. We will not only have to assure we filter out the idioms (as stated in the introduction), but there are also other constructions that are not within our scope of research. Some examples of these will be showed later on in this thesis.

For now, it is interesting to note that I will mainly copy the approach of [Le Bruyn et al., 2009]. Their idea is that the possibility for a bare PP to be able to appear without a determiner is governed by either the preposition or the noun itself. Translating this idea into search queries, one would like to see (1) which prepositions appear frequently with bare nouns and (2) which nouns appear frequently bare after a preposition. In the chapter on English, I will explain in detail how one can accomplish this feat.

1.4 Relation to Artificial Intelligence

As stated on the front page of this thesis, this is a Master's Thesis in (Cognitive) Artificial Intelligence. However, till now, it seems that the topic under discussion is far away from any research in Artificial Intelligence (AI). While I concede that this *seems* to be the case, I am very much obliged to argue that this is certainly not so. In this short section, I will thus link research into aspects of language to the far broader study of AI.

First of all, language is, in my opinion, *the* central aspect of AI. This is not to say that the remainder of AI is mainly bookkeeping, but the omnipresence of communication by language is what makes our species (at least for the moment) something entirely different than any other animal or, as a matter of fact, every artificial piece of intelligence humans have created. This should put language on the list of interests of every self-conscious cognitive scientist.

But of course, such an argument will not suffice to convince the reader. However, there is more. Throughout this thesis, I will employ methods more familiar to AI sympathizers than to those in the linguistic department. These include of course the corpus research described in the previous paragraph (which is, in fact, *data mining*, another central aspect of AI), but also the use of various statistics to back up the claims made. Such a statistics-based approach in this area was first used in this particular field of bare PPs by [van der Beek, 2005], and in this thesis, I hope to provide similar, innovative methods.

Furthermore, as stated some times before now, this thesis also covers research in multilingual corpora, a new tool for linguistic research. Multilingual corpora are most often applied in the context of machine translation. Again, this is more often a subject sought after by researchers in AI than those in the (non-computational) linguistics department. Methods to research such corpora are still under construction⁶, and it is interesting to see whether a computational analysis can extract interesting results from these corpora.

There are therefore reasons enough to label this thesis as dealing with (Cognitive) Artificial Intelligence. Still, I understand that my main public will be linguists. I have thus first and foremost tried to make my thesis readable for them. This means that while linguists will get a decent introduction into the AI-like methods, the AI researchers will get only a minor introduction into the linguist's most favorite topics. However, to help those AI people new in the field of linguistics, in the next section, I will largely summarize the main issues in the land of bare PPs.

1.5 Review of the Literature

In this section, I will reflect on the current state of the research in bare PPs. I will first consider the syntactical research that has been carried out. Since this research has mostly dealt with the 'when' and 'where' of the occurrences of bare PPs, and not very much with the underlying syntactic constructions (with [Baldwin et al., 2006] as a deviation to the rule, as their research focused on accounting for bare PPs within Head-Phrased Structure Grammar), I have dubbed this 'distribution'. In the second part of this section, I will scrutinize the investigation on the semantics of bare PPs. Here, I will also include a short introduction into the fields of prepositions and bare nouns *an sich*, since these might provide some interesting parallels with the bare PPs under scrutiny in this thesis.

1.5.1 Distribution

The Defective NP-hypothesis

One of the first inquiries on both the distribution of bare PPs was carried out by [Stvan, 1998]. She observed that some nouns inside a PP denote not only a location, but also an activity: 'in school', 'at theater' and 'in jail' all denote also to a (prototypical) activity at the specified location. We have already gone over some examples of this phenomenon in the first section of this chapter, but to further illustrate the productiveness of the category, below we find three other examples from Stvan's thesis.

- (1.18) They are not proficient on the computers; like Brian, twenty-four of his classmates have no computers **at home**, and they attend Computer class only twice a week.
- (1.19) While tape recordings to uncover, say, infidelity are not admissible **in court**, they can mean leverage in a settlement.
- (1.20) Up **on deck**, thinking of spending five days on the Dolphin, I began to be seized by feelings of panic and pain I could not explain.

In looking for an explanation as to why these locations would appear bare inside a prepositional phrase, [Baldwin et al., 2006] noted that these nouns also appear without a determiner in argument positions with a similar semantics: in (1.21), (1.22) and (1.23) we find 'school' respectively inside a bare PP, as subject and as object, each time with the meaning 'attending school'. They dubbed these noun phrases therefore **defective** NPs: the nouns in question, denoting a location, seem to appear bare all the time, not only under the licensing of the preposition, as the examples below should show. Baldwin et al. state that we can simply account for these bare PPs by seeing whether they appear bare in subject or object position.

⁶Although some are already well described in [Johansson, 2007].

(1.21) While **at school**, I learned the value of an education.

(1.22) **School** drains the best years of your life.

(1.23) Many students can't afford **school** in the States.

However, there is at least some evidence that the 'defective NP-approach' is neither necessary nor sufficient in explaining which nouns might appear as an object in a bare PP. First of all, there is quite a big deviation in the class of defective NPs: not all locations that appear in bare PPs also appear as either bare subject or bare object. In an earlier work of mine⁷, I used the British National Corpus ([Davies, 2004]) to assess this question.

In this short corpus experiment, I started from a list of locations that appear inside bare PPs from [Stvan, 2007]. For each location in the list, I assessed whether it also occurred bare in subject position and/or object position⁸. Table 1.1 below shows that indeed, some NPs do appear in all argument positions, but this is by no means the rule.

Table 1.1: A typology for defective NPs in English.

	Only in PP	PP and Subject	PP and Object	PP, Subject and Object
Campus	River	Sea	Base	Bed
Cellar	Side		Class	Camp
Country	Slope		Clinic	Chapel
Deck	Stage		Daycare	Church
District	State		Harbor	College
Dock	Stream		Jail	Court
Hall	Studio		Pasture	Home
Hill	Table		Port	Hospital
Island	Temple		Post	Kindergarten
Kitchen	World		Shore	Prison
Market	Yeshiva		Synagogue	School
Meeting			Town	Theater
Planet				University
Property				Work

My conclusion from this table is that there are three classes for Stvan's locations: first of all, there are of course the ones labeled by Baldwin et al. as defective, appearing freely everywhere in a sentence. However, there are also nouns that appear both as a bare object and in a prepositional phrase. Still other nouns seem only appear bare under the governing of a preposition. For a location to appear bare inside a PP, the noun thus not necessarily has to appear bare in argument positions. Therefore, I claim that it is by no means necessary for a noun to be defective to appear in a bare PP.

Concerning sufficiency, we can note that there are a lot of bare PPs which do not follow the defective NP hypothesis. For example, in this thesis, chapter 2, we will see that prepositions like 'per', 'under' and 'by' can take bare NPs as their object which do not appear bare in any other circumstances. Furthermore, in other languages, we do not find that nouns that can appear as

⁷During the Utrecht University course on 'Semantic Structures'.

⁸Here, I queried using the following two regular expressions, which are, in my opinion, quite straightforward, given that [pu*] is used for every kind of punctuation, and [v*] stands for every kind of verb. In the N slot I entered the nouns from Stvan's table.

- For subject usage: [pu*] N [v*]
- For object usage: [v*] N

either a subject or an object in argument positions. Both [van der Beek, 2005] and [Paenen, 2009] note that in Dutch, only 'school' seems to have this property. I thus argue that the defective NP hypothesis is neither necessary nor sufficient in explaining the distribution of bare PPs.

Back to Basics

It might well be that this line of reasoning led to a more 'back-to-basics'-strategy in the field. In these studies, we see that the researchers employ statistical methods on corpora to find interesting types of bare PPs. A prototypical example of such a new, corpus-based strategy is the work of [Le Bruyn et al., 2009]. They extracted all prepositions followed by count nouns from two relatively small corpora (one in French, one in Dutch) and displayed frequency data. From these data points, they were able to quickly find interesting categories of bare PPs in both languages, examples of which will be given in the section on semantics below. They concluded their research with a testable semantic prediction, which will also be reviewed in the next section.

Another example of the new strategy is the research of [Dömges et al., 2007]. Using a big German, computer-annotated corpus they succeeded in proving that the preposition 'unter' ('under') is productive in its combining with bare singular count nouns. This result goes to show that the class of bare PPs deserves extra syntactic attention: now we know that a least one construction is productive, we just can not account for these constructions by claiming them to be idiomatic and belonging to a closed class. Their focus now seems to have shifted from 'unter' to 'ohne', 'without' in English. We will also encounter this latter preposition and review its special status throughout this thesis.

In this thesis, we will employ the strategy proposed by [Le Bruyn et al., 2009]. Using a relatively small corpus consisting of one million words (Brown corpus ([Kucera and Francis, 1967])), we will first of all look for interesting patterns in the bare-PP-landscape in English. After that, we are then able to try and say something about the semantics of these patterns. To be able to do that, it is of course necessary to be aware of previous approaches in this field. I will thus review this literature in the next section.

1.5.2 Semantics

As stated in the introduction of this literary review section, I will now try and summarize some of the research on the semantics side of the bare PPs spectrum. But to do that in a proper way, I think it is necessary to first introduce the two parts of which a bare PP consists, respectively a preposition and a bare singular count noun. While I think that, indeed, it is very important to have a clear view of these before researching bare PPs, in this thesis, due to reasons of time and space, I will only go over some of the main points in this field, giving references here and there for the interested reader. After having introduced prepositions and bare nouns, I will once again focus on bare PPs, and give a proper, more extensive overview of the semantic research.

Prepositions

Prepositions primarily denote position in space. That this is the case, is clear from the way one learns this category in elementary school. In the Netherlands, prepositions are there often denoted as 'kast-woorden' ('closet words'), as one can combine almost each preposition with 'de kast' ('the closet'), yielding a spatial reading. This simple mnemonics is exemplified below in (1.24).

(1.24) voor / achter / vlakbij / op / onder de kast

(1.25) before / behind / near / on / under the closet

For the semanticist, this closed class of (small) words form quite a problem. Dealing with prepositions in a purely logical formalism of predicate logic does not seem to fully capture the

meaning of spatial prepositions. In [Zwarts and Winter, 2000], we thus see a shift to model-theoretic semantics, in which the notion of a vector space is introduced in the type-logical domain.

Vectors are extremely helpful in creating a sensible semantics for prepositional phrases. In a sense, prepositions relate a *figure* (or *located object*) to a *ground* (or *reference object*). For example, if I say that 'John is near the closet', I am stating that the located object John has a position relative to the closet, the reference object. Vectors help to formalize this relationship, as we might in this case that John is located in the vector space that is denoted by the vectors going out of the closet, and having a length that is a small positive number (for 'John is at the closet', this number would be even smaller). Other prepositions could be modeled likewise.

With bare PPs, we see something else happening: the locative part seems to fade a bit. When I tell somebody that 'John is still at school', it does not necessarily mean that he is physically at the school. The phrase only states that he is engaging in scholarly activities (in principle, he could be at home making his homework, or even enjoying his Christmas holidays). We already saw this in example (1.4) above.

With regard to the semantics of bare PPs, it consequently seems that it is not the preposition, but the missing determiner, that is the important factor for acquiring the 'special' semantics. In the following section, we therefore turn to the semantic aspects of bare nominals.

Bare Nominals

The field of bare nominals is broad and therefore not easily summarized in just a few pages. However, some general observations can certainly be made. One of the major starting points formed the research of [Chierchia, 1998]. He claimed after a cross-linguistic research that bare nominals in languages with articles ought not to appear in argument positions. Nevertheless, there is some variation between languages in their usage of articles. The gist of his proposal is that Germanic languages follow Romance in their use of articles for singular arguments (determiners are necessary ("the category D should be projected")), and languages like Chinese for plural arguments (determiners should be left out).

This typology of Chierchia (as he called it, the 'Nominal Mapping Parameter') might well have led researchers to find counter examples of bare singular count nouns in argument positions. In earlier parts of this thesis, we for example have seen that [Stvan, 1998] found that nouns of the LOCATIONS-category ('school', 'church', 'prison' et cetera) might appear bare in argument positions. Here, we saw that these locations acquire an enriched semantics when used bare: 'school was nice' refers to scholarly activities, not solely the location.

Stvan was not the only one to search for bare arguments: e.g. [Borthen, 2003] searched and found bare singular count nouns in argument position in Norwegian. Borthen defines four 'licensing situations' of bare nominals in argument positions. Interestingly, in her thesis, Borthen also gives some examples in which the licensing situation also occurs with bare PPs. I will focus on these licensers in the now following paragraphs.

The first licenser of bare singulars is *conventional situations*: if the situation described is conventional, the noun can appear bare. Borthen shows so in the first two examples below: while (1.26) is grammatical in Norwegian (as sleeping in a hammock is a conventional situation), (1.27) is infelicitous (one normally does not sing in a hammock). A second licenser would be so-called *have-predicates*: after 'with' and 'without', bare nominals can serve as a valid argument, as in (1.28) and (1.29). We will see later on that these prepositions will return in both the English and Dutch analysis.

(1.26) *Per sov i hengekøye.*
Per slept in hammock.
"Per slept in a hammock."

(1.27) * *Per synger i hengekøye.*
Per sings in hammock.

"Per is singing in a hammock."

(1.28) *Jeg fikk et brev uten frimerke.*
I got a letter without stamp

"I got a letter without a stamp."

(1.29) *Dette er en bil med stor motor.*
This is a car with big motor

"This is a car with a big motor."

The other two licensors of bare singulars in Norwegian (*comparison of types* and *covert infinitival clauses*) do not seem to have anything to do with bare PPs, and we will thus not work out the details of these licensors here. Having with both Borthen and Stvan linked the subject of bare nominals to bare PPs, in the next section, we will look at research from authors having bare PPs as their main focus.

Bare Prepositional Phrases

Concerning the semantics of bare PPs, we should again start off with research from Laurel Stvan. In both her PhD thesis ([Stvan, 1998]) as in her later articles ([Stvan, 2007] and [Stvan, 2009]), she gave insight in the semantics of locational bare PPs like 'in prison', 'on campus' or 'at school'. She did this on the bases of gathered examples from newspaper articles. We saw in the introduction that one of the conclusions was that these bare PPs denote not only a location, but also the typical activity that is usually performed in the location. For example: 'John is in prison' means not only that John is in a prison, but also that John is serving time as a prisoner. A semantics for these bare PPs should of course take these observations into consideration.

Her main conclusion however is not only that there is semantic enrichment. In [Stvan, 2007], she discerns three readings of bare singular count nouns in prepositional phrases. First of all, there is the **activity** reading just mentioned in the previous paragraph, exemplified again below in (1.30) below.⁹ Here, we see that these companies have skirmished not only in the courthouse, but also especially during trial proceedings. However, Stvan also notes that bare PPs can render a **familiarity** reading. In (1.31) below, for example, we see that 'in town' denotes the speaker's town. The last reading she finds among her data, is the **generic** reading in (1.32) Here, with the help of the bare PP 'on campus', something is said about campuses in general. One could easily replace the singular noun with its plural, yielding the same meaning.

(1.30) Off and since then, the companies have skirmished **in court**. #They had a great time there.

(1.31) My dad was **in town** the weekend before my birthday. He had a great time there.

(1.32) "Free speech", "Question authority", and "Leave us alone" are now conservative and libertarian battle-cries **on campus**. *Because of that, I had a great time there.

As is clear from the continuations I have added to Stvan's examples, the three readings differ in their setting up of discourse referents. While the familiarity reading sets up an easily picked up referent, with the activity reading, this seems to be more of a problem. And with the generic reading, it is clear no referent is introduced. Stvan hypothesizes that, in other languages, one might see that this different meanings are rendered by different forms. However, in the examples she gives, we only find a disambiguation of the activity reading and the purely locative sense (e.g. French uses 'en prison' for the activity reading, and 'dans la prison' for the regular locative construction).

While these different senses are certainly helpful in creating the fitting semantics for these bare PPs, one important question remains unanswered: how the *absence* of form can lead to an *enrichment* in meaning? [Stvan, 2009] argues that **semantic incorporation** is the answer. She

⁹The examples here are from [Stvan, 2007].

explains that these locational bare PPs are in many ways similar to incorporated nominals in, for example, Hungarian ([Farkas and de Swart, 2003]). In Hungarian, bare nominals only occur in specific expressions and then also has some specific properties, which I have listed below:

- *Discourse Transparency*: the expression does not introduce a discourse referent.
- *Lack of Modification*: the expression cannot be modified without losing grammaticality or its alternative semantics.
- *Number Neutrality*: the expression does not place any restriction on the number of referents involved.

From some examples that I have mentioned already in the introduction, it is clear that bare PPs have these properties. For example, in (1.4) above, repeated below in (1.33), we saw that bare PPs do not introduce a discourse referent. Secondly, modification often leads to ungrammaticality of the bare PP, shown in (1.6) above and (1.34) below. And lastly, there sometimes is number neutrality with bare PPs, as was shown in (1.32) above, displayed again in (1.35).

(1.33) *Discourse Transparency*: Pat is in prison. #It is a 3-story concrete building.

(1.34) *Lack of Modification*: Pat is in *big / *red / federal / state prison.

(1.35) *Number Neutrality*: In September, she pleaded guilty and paid a \$500 fine. Her alternative was 90 days in jail. And you know, **this** / **these** can be cruel nowadays.

While the observation that bare PPs are strikingly similar to Hungarian bare nominals is indeed very interesting, Stvan does not seem to work out the exact semantics, for example within Discourse Representation Theory. It is therefore not entirely clear whether semantic incorporation is the right way to account for the differences in semantics between full and bare forms. No, here, as [de Swart and Zwarts, 2009a] argue, one finds a typical case in which Optimality Theory can save the day. Basically, in semantics¹⁰, Optimality Theory (OT) states that marked forms come with marked meanings, and unmarked forms yield an unmarked meaning. However, with bare PPs, we see that the forms with more semantic weight are syntactically less marked.

How can less than be more? De Swart and Zwarts argue that in principle, this phenomenon is not that strange. They claim that in other constructions, one also finds this division of semantic labor. And in the case of bare PPs, the account is actually quite straightforward. With bare and full PPs, we can argue that the latter form is marked, as its having a determiner makes it a more complex form. And concerning semantics, we might argue that with 'in prison' vs. 'in the prison' a reading in which someone incarcerated has less semantic weight compared to a reading in which someone is only in the prison on visit, as the former reading is regarded as more prototypical (one normally is in a prison to serve a sentence, not to visit a relative (though this might be more clear with 'school')).

Filling in the OT tableau with these two constraints (*F for form and *M for meaning) defined and our four possible options would then lead to Table 1.2 below. The proper (and of course wished-for) mappings are here then the outcome of bidirectional optimization.

However, this short story seems to be a bit too short. As we have seen before, the bare form is actually not the standard form: under normal circumstances, bare nouns in argument position are clearly out. The full form is the mode and would then be the unmarked form. Concerning the semantics, we have a different problem. How should we argue that "imprisoned" is actually less marked than "just visiting"? There is no easily observable difference in complexity of meaning here.

¹⁰For an introduction into Optimality Theory in semantics, I refer the reader to [Blutner, 2000].

Table 1.2: Bidirectional optimization for bare PPs, first version.

			*F	*M
a.	☞	in jail, ‘imprisoned’	✓	✓
b.		in the jail, ‘imprisoned’	*!	✓
c.		in jail, ‘just visiting’	✓	*!
d.	☞	in the jail, ‘just visiting’	*!	*

For these two problems two disappear, [de Swart and Zwarts, 2009a] define two new constraints which should make the tableau more perspicuous. For form, they define the constraint *FUNCTN, which simply states to avoid functional structure in the nominal domain. It is clear that the full PP violates this constraint, while the bare PP does not, as there is no overt article. With regard to the semantics, we might want to write out the two logical semantics for either being incarcerated or being only on visit:

$$(1.36) \lambda x.(in(x, y) \wedge jail(y) \wedge imprisoned(y, x))$$

$$(1.37) \lambda x.(in(x, y) \wedge jail(y))$$

We see here that (1.36) is a proper subset of (1.37). We could thus argue that the former is a stronger meaning than the latter. The constraint STRENGTH (defined as “stronger meanings are preferred over weaker meanings”) would formalize this relationship. If we then again go over the four possible relations and the two constraints, we end up with the following tableau in 1.3, in which it is totally clear how the absence of form leads to an enriched meaning.

Table 1.3: Bidirectional optimization for bare PPs, second version.

			*FUNCTN	STRENGTH
a.	☞	in jail, $\lambda x.(in(x, y) \wedge jail(y) \wedge imprisoned(y, x))$	✓	✓
b.		in the jail, $\lambda x.(in(x, y) \wedge jail(y) \wedge imprisoned(y, x))$	*!	✓
c.		in jail, $\lambda x.(in(x, y) \wedge jail(y))$	✓	*!
d.	☞	in the jail, $\lambda x.(in(x, y) \wedge jail(y))$	*!	*

We now have a workable OT semantics for a special kind of bare PPs. But as for example [Baldwin et al., 2006] and [Le Bruyn et al., 2009] note, there are other kinds of bare PPs to be found. Both articles mention a whole range of bare PPs that do *not* easily fit within the story of semantic enrichment. An example is the preposition ‘per’ in Dutch (mostly translated by ‘per’ in English, but in chapter 5 we will see that this certainly is not always the case). This preposition seems to open a slot where every noun could fit in, yielding a reading in which the noun is strongly individuated. I have copied some examples from Le Bruyn et al. below:¹¹

$$(1.38) \text{per: per regio (per region), per manuur (per man-hour)}$$

¹¹While Baldwin et al. might have been the first to make a clear distinction between the various types of bare PPs, in the now following paragraphs, I will mainly stick to the analysis of Le Bruyn et al., as they seem to have a more worked-out approach on the differences between various kinds of bare PPs.

Here, the licensing of the bare occurrence of the noun inside the prepositional phrase is certainly not due to the status of the noun, but far more likely due to the status of the preposition. The preposition opens a slot in which any noun can fit in. Most of the time, however, this slot has a lexical restriction. The slot is then only to be filled in by a specific category of nouns, sharing similar properties. An example is ‘en’ in French.

(1.39) en SHAPE: en amande (almond shaped), en colimaçon (snail shaped)

There is consequently ample evidence to make a distinction between these two forms. First of all, there exist bare PPs in which the noun licenses the bare occurrence, and in which there is competition in both form and meaning (the ‘at school’ and ‘in prison’ cases, in which the full PP is grammatical and yields a weaker semantics). Le Bruyn et al. call these **N-based bare PPs**. Next, there are bare PPs in which the preposition is the licenser of the bare PP. These bare PPs have no competition in form and meaning, since the full PP would be regarded ungrammatical in nearly all cases. The logical name for this kind of bare PPs would then be **P-based bare PPs**.

To provide a proper semantics for these two kinds, Le Bruyn et al. start from a general semantics for any preposition P , a figure u and a ground x : $\lambda x.\lambda u.P(u, x)$. For the N-based bare PPs we find that the semantics comes down to (1.40) below, for P-based bare PPs the semantics will be like (1.41).

(1.40) $\lambda N_{lex}.\lambda u.\exists!x[N_{lex}(x) \wedge P(u, x) \wedge Poss(u, x) \wedge A(u, x)]$

(1.41) $\lambda N_P.\lambda u.WDx[N_P(x) \wedge P(u, x)]$

Here, (1.40) puts some constraints on the semantics. First of all, it states that the noun in question belongs to a special lexical class ($N_{lex}^{e \rightarrow t}$); e.g. the locations like ‘prison’ and ‘school’. On the other hand, the choice of the preposition P is relatively free. Secondly, there is a possessive $Poss$ relation between the figure and the ground: when we say that ‘John is at school’, John is learning at John’s school, not at Mary’s, if she happens to be educated at a different school. The activity sense of the bare PP is denoted by the A -relation between figure and ground. Lastly, the noun in question should be one of its kind ($\exists!x$). With this latter constraint, we see that [Le Bruyn et al., 2009] are only interested in the activity and familiarity sense defined by [Stvan, 2007] below, and certainly not in the generic sense in (1.32) above, where the singularity of the noun is of no important. It might well be that this generic usage needs a separate analysis.

As for the P-based bare PPs, the semantics is yet very sketchy. It now states that indeed, the nouns that are to be filled in are defined by the preposition ($N_P^{e \rightarrow t}$), and the preposition P again is relatively free. Furthermore, we see that the bare PP as a whole gets a weak definite reading, as the bare noun in question can not be referred to in the next sentence.

Can we find even other categories of bare PPs? Le Bruyn et al. answer this question positively. They find that in both Dutch and French, the prepositions that are the translations of ‘with’ and ‘without’ (respectively ‘met’ and ‘zonder’ in Dutch, and ‘avec’ and ‘sans’ in French) are special. Here, the meaning of the bare PP is not in any way different from that of the full PP: the utterances ‘de man met hoed’ (‘the man with hat’) and ‘de man met de hoed’ (‘the man with the hat’) are essentially conveying the same meaning: that there is a man that has a hat. There thus is some competition in form, but not in meaning. Le Bruyn et al. call this kind of bare PPs **\exists -based bare PPs**, as these prepositions here allow bare nouns to get an indefinite interpretation. The semantics of \exists -based bare PPs should be something like (1.42).

(1.42) $\lambda N.\lambda u.\exists x[N(x) \wedge P_{with}(u, x)]$

For these \exists -based bare PPs, there is no constraint on the noun. The semantics of the preposition P_{with} is left somewhat implicit. The expression does specify an existential relation.

For the last type of bare PPs, **D-based bare PPs**, Le Bruyn et al. turn to Romanian. As [Mardale, 2006] proves, in Romanian, a non modified noun preceded by a preposition is necessarily used *without* the definite article, but the interpretation is actually the definite one. We

thus find competition in form, but not in meaning, and since the bare form is less marked, the bare alternative wins. When the noun preceded by the preposition is then modified, the definite article is required, which we also found with N-based bare PPs. The semantics of D-based bare PPs is quite straightforward (and the choice of N is free, just as with \exists -based bare PPs above):

$$(1.43) \lambda N.\lambda u.\exists!x[N(x) \wedge P(u, x)]$$

Having given an overview of the literature on bare PPs, we are now in a position to sharpen the research questions. As might have been clear by the extensive review of [Le Bruyn et al., 2009] in the last paragraphs, in this thesis, we will mainly follow the typology that Le Bruyn et al. set up for bare PPs. That is, in each of the two monolingual chapters (on English and Dutch), we will look for P-based, N-based and \exists -based bare PPs. After these two chapters we are then able to do a proper comparison on these three kinds of bare PPs. It is obvious that then, some questions will still be left unanswered. It is these questions that we will try to answer in the multilingual part.

1.6 Outline of this Thesis

This thesis is divided in two parts. In the first part, we will look at bare PPs from a monolingual perspective. In the first chapter, we will try to say something about the distribution of bare PPs in English. We will employ the method of [Le Bruyn et al., 2009] to find interesting pairs of prepositions and bare nouns in the Brown Corpus ([Kucera and Francis, 1967], freely distributed within the NLTK package [Bird et al., 2009]). Also, earlier results from [Stvan, 1998] are reported and attested. The second chapter will deal with bare PPs in Dutch. We will review the data from both [Paenen, 2009] and [Le Bruyn et al., 2009], but also include some new observations using the Eindhoven Corpus ([van Grootheest, 1989]).

In the second part of this thesis, we will see whether a multilingual perspective on bare PPs might any new light on the monolingual analysis. Throughout this part, we will use a parallel corpus of European Union proceedings ([Koehn, 2005]) to find relations between the languages for the languages for which an individual analysis is present. In the first chapter, we will try to prove some larger hypotheses in the domain of bare PPs using a statistical analysis. In a second chapter, we will opt for a more in-detail analysis, per preposition. Here, the data that we gathered in the first part of the thesis will certainly come in handy, but we will also see that the multilingual corpus helps us define more clearly both the senses and the semantics of prepositions in combination with bare singular count nouns.

At the end of this thesis, I will include a chapter on both the advantages and technical problems of multilingual corpora, making use of a multilingual corpus I compiled myself. After that, this thesis then ends with a conclusion, in which the main results are reported and suggestions for further research will be given.

Part I

The Monolingual Perspective

Chapter 2

Evidence from English

2.1 Introduction

In this chapter, I will search for bare PPs in the English language. While there has already been quite a lot of research on bare PPs in English, I believe that a new analysis is warranted for reasons explained in the next paragraph. After giving this justification, I will present my method of research for finding bare PPs. I will shortly return to the various types of bare PPs sought after. Also, some drawbacks of the proposed method will be considered. When the methodological section is finished, I will show the results and end this chapter with a short conclusion.

2.1.1 Why Again English?

We saw in the introduction of this thesis that both [Stvan, 1998] and [Baldwin et al., 2006] have given us some direction of where to search for interesting patterns in the field of bare PPs. I would like to argue, however, that both researchers might have had a wrong focus: while Stvan was focusing on bare PPs denoting an activity instead of a location (at church, in prison, during school), and carefully worked out both the distribution (in [Stvan, 2007]) and the semantics ([Stvan, 2009]) of this class, she might have missed some other constructions due to her somewhat narrow perspective. And when Baldwin et al. set out on a journey to find a good way to account for bare PPs within Head-Phrase Structure Grammar, their focus might have been too broad, and they might also have overlooked some productive constructions on their way. Consequently, there is more than enough reason to take another, structured and in-detail look at the English data.

2.1.2 Method

For this more in-detail look at the English data, I employed the method of [Le Bruyn et al., 2009]. They used relatively small corpora (around one million words) to find interesting classes of bare PPs in both Dutch and French. In English, we find a similar corpus in the Brown Corpus [Kucera and Francis, 1967]. This corpus consists of 1.014.312 words of English prose, all written by native speakers of American English. The corpus contains both fiction and non-fiction, and, importantly for our purposes, is tagged for part of speech.

The method that [Le Bruyn et al., 2009] use to extract bare PPs from a corpus is simple, yet effective. They scan the corpus for all occurrences of prepositions directly followed by a singular noun. For the Brown corpus, this means that the search query (a so-called *regular expression*) is as follows:

(2.1) `'[a-z]*/in [a-z]*/nn'`

So, we search for any ('[a-z]*' is the regular expression for zero or more letters between a and z, i.e. words) preposition (marked by the tag '/in' in the Brown Corpus) followed by any (again '[a-z]*') singular noun ('/nn'). Also note the extra space on the end of the query, which excludes, for example, possessive singular nouns (which are tagged '/nn\$ '). Each hit is then stored in a database table with two columns, [prep *varchar(100)*] and [noun *varchar(100)*]. Using Structured Query Language (SQL) we are able to group the data in any way we like.

2.1.3 Drawbacks to the Method

An important drawback to this approach is that while the search query in (2.1) does exclude some uninteresting patterns from the data, some are still included and pollute the data. In the examples below, we find all kinds of patterns which are not excluded using the search query in (2.1) above. This is an important thing to remember when one tries to make any statistics on the data. I will assume that this 'pollution' occurs uniformly over each preposition, so that the provided numbers are only off in a certain percentage.¹

(2.2) [P N N] Of course the factor **of head start** made all the difference.

(2.3) [P N P] **In response to** your letter, I can in good conscience recommend my son, David, in the field of leadership.

(2.4) [P N P N] We live **from crisis to crisis**.

(2.5) [N P N] **Day after day** some new episode is reported.

(2.6) [P N ^ N] An enthusiastic audience confirmed the "live" character of the hour, and provided the interaction **between musician and hearer** which almost always seems to improve the quality of performance.

Another serious drawback to the current approach is that it does not make any distinction with regard to mass and count nouns. Since the corpus is not tagged for this distinction, we will find both mass and count nouns in our statistics, while strictly spoken, we are only looking for countable nouns. Manual analysis is therefore required to filter out the mass nouns, only to be left over with the bare singular counts nouns we are after.

2.1.4 Two kinds of bare PPs

In their paper, [Le Bruyn et al., 2009] make a distinction between two kinds of bare PPs. They first of all search for **P-based bare PPs**. In this kind, the bare occurrence of the noun is licensed by the preposition governing it. They hypothesize that the preposition opens a slot in which nouns with the right lexical semantics might fit in. Le Bruyn et al. found some patterns for Dutch and French. Two examples (one for each language) are provided below in (2.7) and (2.8). Here, the idea is that the preposition opens a slot in which every noun of the lexical category given should fit. The obvious question now is of course: can we find the same kind of P-based bare PPs in English?

(2.7) per MEANS OF TRANSPORTATION: per ballon (by balloon), per auto (by car)

(2.8) en SHAPE: en amande (almond shaped), en colimaçon (snail shaped)

Within P-based bare PPs, 'with' and 'without' form an interesting deviation from the norm, as here there is less semantic effect from the leaving out of the determiner. In Le Bruyn et al., 'with'

¹I am aware that this assumption could be far off the truth, leading to a wrong image of the world of bare PPs. However, as I will note again and again, the frequency data that will be presented are only used to find the most promising prepositions, and the numbers presented in this chapter are by no means used/useful to draw statistical conclusions.

and ‘without’ thus form a distinct class of bare PPs; the \exists -based bare PPs. In this thesis, I will follow this suggestion and scrutinize these prepositions in a separate section.

The logical complement of the P-based bare PPs form the **N-based bare PPs**. Within this class, we find nouns that appear among a multitude of prepositions. In these cases, the noun seems to govern its own bare appearance. In this class we find, for example, the locations and interruptions of [Stvan, 1998]. These nouns appear with several distinct prepositions. An example can be found below in (2.9). Also, in this class, we find the pragmatic enrichment Stvan first noticed. Interestingly, [Le Bruyn et al., 2009] did not find a lot of these nouns in Dutch nor French.

(2.9) at/in/from/outside/to/toward school

In the following two sections, we will first of all search our corpus for P-based bare PPs. After that, we will turn to \exists -based bare PPs. In the last part of this chapter, we will shift to our focus to N-based bare PPs.

2.2 P-based bare PPs

As explained in the previous section, P-based bare PPs are essentially prepositions which open a slot for a noun of a specific lexical category to appear bare. How are we going to find these? What we could do, copying [Le Bruyn et al., 2009], is to first of all make a frequency table, in which we state the amount of bare nouns per preposition. If this amount is high with respect to the amount of times the preposition occurs in the corpus, it might be that we have stumbled upon a P-based bare PP. A closer analysis should then help us decide whether this is indeed the case.

So, the method is as follows. First of all, count for each preposition P_i the number times it occurs directly followed by a bare noun. For each preposition P_i , this number is divided by the total numbers of occurrences of P_i in the corpus. Doing so, one obtains for each P_i a percentage of bare nouns it occurs with. Ordering these relative frequencies then might lead us to interesting preposition-noun-combinations.

To copy this procedure, I extracted all preposition-noun-combinations from the Brown Corpus into a database table. Then, I used a SQL query to model the procedure provided in the previous paragraph. In Table 2.1, we then find the results of this procedure. A graphical representation of this table can be found in Figure 2.1 below.

Note again that this is quite a rough procedure to extract the data from the corpus. Since one only searches for $[P N_{singular}]$ patterns, one might also include all kind of unwanted patterns into the statistics. The number and percentages are therefore only indicative: one should not try to draw any statistically significant conclusion from the data in Table 2.1 and the subsequent graph.

On the other hand, this preprocessing of the data *does* help to direct us towards interesting bare PPs. As [Le Bruyn et al., 2009] note, in Dutch, we also find that ‘per’ is most likely to be followed by a bare nominal. In English, this is also the case. Do we find a similar distribution of nouns there? Another interesting result is the distinction between ‘with’ and ‘without’. Why is it that these antonyms are so different in their percentage of occurrences with bare nouns? In the next section, I will try and answer these research questions one by one, by taking a closer look at some of the prepositions in the table.

The research in each of the following sections proceeded by the following steps:

1. Using a SQL query to select the nouns occurring with this preposition and their count from the database table.
2. Disregarding all clear idioms (recall the discussion we had on idioms in the introduction of this thesis). These idioms are recognizable by
 - often a high number of occurrences and
 - appearance in a dictionary.

Table 2.1: Frequency data for English. From left to right: preposition, number of occurrences, number of bare nouns, percentage of bare nouns. In the penultimate column, we find the noun the preposition occurs most common (m.c.) with. For reasons of space, I have chosen to omit prepositions with less than 50 total occurrences in the corpus.

P	#	# BN	% BN	m.c. BN	#
per	370	323	87,30%	cent	146
without	574	139	24,22%	regard	5
concerning	62	13	20,97%	effectiveness	1
of	35023	6722	19,19%	course	324
under	686	115	16,76%	way	17
for	8843	1398	15,81%	example	173
in	20724	3175	15,32%	fact	154
into	1778	237	13,33%	account	19
after	698	93	13,32%	independence	5
between	725	93	12,83%	onset	3
before	406	52	12,81%	dawn	6
with	7260	851	11,72%	respect	66
including	166	19	11,45%	mountain	1
by	5220	583	11,17%	means	20
but	131	14	10,69%	trot	2
upon	449	42	9,35%	completion	2
on	6103	562	9,21%	top	30
off	156	14	8,97%	balance	6
from	4340	388	8,94%	time	27
toward	384	34	8,85%	area	2
to	11044	977	8,85%	death	28
through	905	76	8,40%	faith	3
beyond	162	13	8,02%	recognition	2
towards	63	5	7,94%	town	1
against	622	49	7,88%	distance	2
below	80	6	7,50%	ground	3
at	5352	380	7,10%	night	46
during	583	40	6,86%	adolescence	7
past	61	4	6,56%	midnight	2
near	156	10	6,41%	unanimity	1
inside	79	4	5,06%	holder	2
outside	83	4	4,82%	government	1
about	1236	58	4,69%	religion	2
above	188	8	4,26%	ground	2
throughout	133	5	3,76%	history	4
around	326	12	3,68%	town	1
than	496	17	3,43%	average	1
until	124	4	3,23%	dinnertime	1
behind	229	7	3,06%	glass	2
among	369	11	2,98%	motor	1
within	351	9	2,56%	control	1
over	835	21	2,51%	time	2
except	173	4	2,31%	knowledge	1
along	198	4	2,02%	edge	1
across	255	5	1,96%	party	1
since	180	3	1,67%	dawn	1
despite	104	1	0,96%	section	1
as	121	1	0,83%	director	1
up	176	1	0,57%	front	1

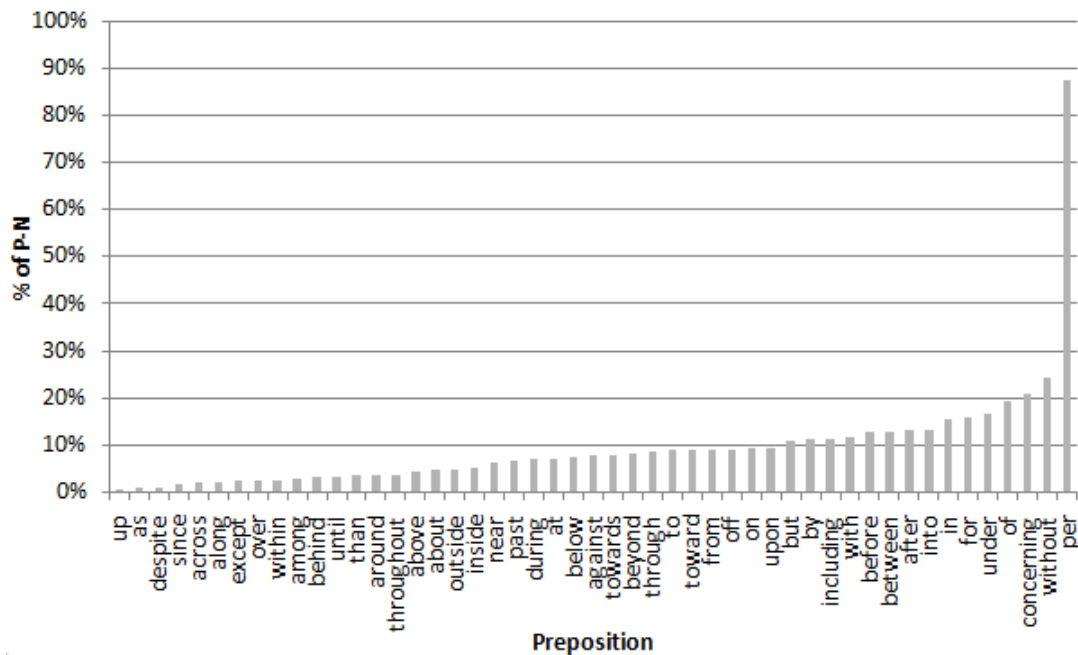


Figure 2.1: Relative frequency of bare nouns per preposition.

- Seeing whether the extracted nouns were countable. This countability check was performed by exploiting a method of Tibor Kiss (University of Bochum), which has yet stayed unpublished. For a noun N to be countable, Kiss states that it has to satisfy four conditions:

"A has more N_{sg} than B" is rendered ungrammatical. (e.g. * A has more **question** than B)

"A has more N_{pl} than B" is rendered grammatical. (e.g. A has more **questions** than B)

"A has more types of N_{sg} than B" is rendered ungrammatical. (e.g. * A has more types of **question** than B)

When giving a definition of the noun, the indefinite article is obligatory. (e.g. (* A) question is an illocutionary act that has a directive illocutionary point of attempting to get the addressee to supply information.²)

- Trying to categorize the remaining bare singular count nouns by hand. For example, the occurrences of 'bus', 'train' and 'car' can be categorized as MEANS OF TRANSPORTATION. Note that there were no categories defined beforehand. There might well be a possibility that another researcher might have extracted other categories from the data.

In the now following section, I will report the results per preposition. I will start off with the highest preposition in the table when ordered on percentage of bare nouns, 'per', and then work down the table, skipping some of the prepositions for which I expect a large amount of either idioms (e.g. with 'of' and 'for') or mass nouns following the preposition (e.g. with 'concerning').

²Definition from <http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsAQuestion.htm>

2.2.1 *Per*: a UNITer

In English, as in Dutch (see [Le Bruyn et al., 2009] and the following chapter), we find that 'per' combines most easily with bare nouns. Even stronger: in both languages we do not find a single example in which 'per' is followed by a determiner. The preposition is consequently in both languages on top of the frequency table. Some examples of 'per + N' can be found below:

(2.10) per day, per pound, per mile, per acre.

(2.11) per head, per person, per patient, per student.

(2.12) per cent, per thousand.

In both (2.10) and (2.11), we find that the construction as a whole gives rise to a UNIT reading: 'per + N' can be paraphrased as 'for each N'. An small deviation from the norm can be found in (2.12): here we see that a 'number' follows 'per', yielding a similar reading to the other examples (50 per cent = 50 for each 100, 50 per thousand = 50 for each 1000).

From these examples, it does not seem that 'per' is *that* interesting. However, there is something special about 'per'. First of all, 'per' does not have a spatial reading like other prepositions. Secondly, and maybe because of that, its semantics is quite unlike that of other prepositions. And lastly, in Dutch and French, 'per'/'par' can also combine with modes of transport or communication. Therefore, in chapter 5, we will try to work the details of 'per' with the help of a multilingual comparison with Dutch.

2.2.2 *Under* under study

For 'under', there is more than one reason to give the preposition a closer look. Of course, first of all, 'under' is ranked high with respect to the relative frequency of bare nouns following it. But secondly, [Dömges et al., 2007] devoted a whole article to the preposition: they managed to prove that the German equivalent of 'under' ('unter') is productive. Is this also the case in English? In this paragraph, I would like to show that indeed, 'under + N' is a productive construction.

For 'under', we again follow the main strategy outlined in the introduction of this chapter. In the Brown Corpus, we find that the nouns following 'under' and at least twice in the corpus can be divided into the three categories, where the last category in fact consists of nouns which can not be easily placed within one of the former two. Between brackets we find the number of occurrences in the corpus.

(2.13) under SECTION: section (2), subsection (2)

(2.14) under ACTION: control (6), consideration (4), discussion (4), study (4), contract (3), refrigeration (3), review (3), advisement (2), arrest (2), authority (2), construction (2), investigation (2), scrutiny (2)

(2.15) exceptions: way (17), fire (6), pressure (4), par (3), water (3), equilibrium (2), pretence (2), vacuum (2)

The 'under SECTION'-category seems to be less interesting for our present purposes. Each construction in this category is followed by a number or literal (or both) of where one can find the subject ("One can read more about this under section 42c"). The nomina in this category *always* appear bare (also with other prepositions or in argument positions), and are therefore not of interest. The construction as a whole seems to be a close relative of constructions like 'under Queen Elizabeth' or 'under coach Martin', which are excluded from the data beforehand.³

But the category in (2.14) is of major interest, as this might be a productive category. Let us take a closer look at some of the examples:

³In the Brown Corpus, these constructions are tagged *nm-tl*, and are consequently not extracted when one employs the search query in (2.1).

(2.16) While this was **under consideration**, dauntless as ever Wright set about the building of Taliesin 3.

(2.17) The resolution **under discussion** at the convention was to require the boards of election to instruct judges to properly display the American flag.

(2.18) City Finance Director Richard J. McConnell endorsed the higher fees, which, he said, had been **under study** for more than a year.

Take note that here, in contrast with the earlier examples of 'at (the) school' and 'in (the) prison', but in line with the examples with 'per' above, we do *not* find any competition in form and meaning with the full PP, as forms like 'under the consideration' are ungrammatical. However, in line with Stvan's locations, but in contrast with 'per', with 'under', there is only a limited possibility of modification.

In the corpus, we find 29 types (yielding 60 tokens, which is 52% of the total tokens) of 'action nouns' as we dubbed them above. To see whether this category is productive, we might opt for the Semantic Substitution test defined in the introduction of this thesis. Indeed, we find several semantically related nouns appearing after 'under' (e.g. 'under attack/siege', 'under development/construction' and 'under consideration/discussion/study'), and thus, the Semantic Substitution test is passed. Furthermore, when doing a Google Search for 'under' followed by any noun, we indeed find the familiar 'under construction', but also 'under destruction'. This kind of wordplay also goes to show, in my opinion, that the class is indeed productive.

So, we have found yet another productive class of bare PPs. But what are then the semantics of the 'under ACTION'-category? I would like to argue that these take over the role of the passive. To prove this hypothesis, I provide the minimal examples in (2.19), (2.20) and (2.21) below (checked for grammaticality with Google).

(2.19) This was **under consideration** ⇔ This was **being considered**

(2.20) The resolution **under discussion** ⇔ The resolution **being discussed**

(2.21) The fees had been **under study** ⇔ The fees had been **studied**

As is clear from the examples, I suggest that for the bare PPs in (2.14) every 'under + N' (with N of the ACTION category) can be substituted *salva veritate* with 'being N-(e)d'. This would say that the 'under + N' construction is, in fact, an alternative for the passive, in which the subject has vanished. It is even stronger: when one uses 'under', the subject can not be added to the sentence, as (2.22) below shows (checked for ungrammaticality with Google).

(2.22) A thesis is under discussion (# by a professor)

So, I propose that the semantics for 'under' would be that the preposition picks up the subject of the sentence and turns it into an existential quantifier (there is someone who discusses the thesis, but who it is exactly, is and can not be specified). In this manner, this preposition seems quite unique: it totally loses its locative meaning which it would have with other, determined compliments (e.g. 'under the bridge').⁴

One last note on the distribution: from the previous, one would expect that every nominalization of a transitive verb can appear after 'under', yielding a passive reading. But this seems not to be the case. Not every nominalization of transitive verb can be readily positioned after 'under'. Utterances like 'under kill' or 'under love' are, for example, ungrammatical. Therefore, while this 'nominalization principle' seems to be a necessary condition, it is by no means sufficient for explaining the distribution.

⁴Still, one could argue that 'under' followed by a bare nominal still has something 'hierarchical' in it: when a certain entity e_1 is 'under study', there is some entity e_2 that is performing the studying, and in a way e_2 can then be regarded as above e_1 . However, I will leave these details of the semantic analysis for other researchers to explore.

2.2.3 *In* and *By*: a pair of antonyms?

As we have seen and will see throughout this thesis, the pair 'with' and 'without' form an interesting antonym. But what about other prepositions? While 'by' and 'in' are by no means directly related, in their choice for bare nouns, they do seem to have some uncommonalities. Let us scrutinize these after having looked at both prepositions one after the other. First of all, we will analyze 'by', which was found to be the translation of some types of 'per'-usage in Dutch. For 'in', there is also a Dutch connection: [Paenen, 2009] found that 'in' in Dutch freely combines with almost every noun which denotes a piece of clothing. Of course, it is interesting to see whether English has a similar arrangement for this category of nouns.

But let us start with 'by'. The preposition 'by' licenses (at least⁵) four categories of nouns:

(2.23) by MEANS OF TRANSPORTATION: by train, by air(line), by (steam)boat, by automobile, by bus, by car, by space(craft), by subway, by truck, by rail, by road, by tramway.

(2.24) by MEDIUM: by (tele)phone, by letter, by radio, by (air)mail, by book, by postcard.

(2.25) by SHORT PERIOD OF TIME: by day, by moonlight, by night(fall), by (after)noon, by sundown, by dawn, by midnight.

The class in (2.23) also appears in Dutch, the MEANS OF TRANSPORTATION class follows bare after 'per' or as a full DP after 'met' in the same (weak definite) meaning. In English, we find that 'by' fulfills this job, as we have also seen in the previous section on 'per'.

The MEDIA in (2.24) also appear in Dutch, then often preceded by 'via' (though not always with a bare noun, e.g. 'by letter' ⇒ 'via/met een brief'). In a sense, these MEDIA are related to the MEANS OF TRANSPORTATION, as both are used for transporting: in the first case, a message is transported, while in the second, its mostly about people being transported.

Note that not all of the nouns in the category numbered by (2.25) are strictly countable. Still, they form interesting categories, as can be seen when one categorizes the nouns that appear bare after 'in'. The nouns following 'in' can be divided into five categories.

(2.26) in CLOTHING: in uniform, in shirt.

(2.27) in LOCATION: in town, in school, in bed, in space, in college, in prison, in court.

(2.28) in MEDIUM: in radio, in film, in television.

(2.29) in PROLONGED PERIOD OF TIME: in spring/summer/fall/winter, in August/September, in daylight.

(2.30) in PROLONGED PERIOD IN LIFE: in life, in marriage, in death, in childhood, in adulthood.

The examples in (2.27) and (2.28) are somewhat equivalent in meaning. In both cases the noun does not denote the actual location. When you are 'in prison', you are not just visiting, but you are incarcerated. The same idea is applicable to the examples in (2.28) (e.g. "I was in radio from 1943 to 1974"), although the alternative meaning of being 'in the medium' is absurd. These differences in meaning of the examples in (2.27) are widely documented in the literature on bare nouns (cf. for example [Stvan, 1998], [de Swart and Zwarts, 2009a] and the introduction of this thesis), but as far as I am aware, (2.28) has not yet been mentioned.

However, this latter category does not seem to be really productive. For most other nouns in the MEDIUM-category (e.g. 'telephone' or 'newspaper') we will not find such a use. Such is not unlikely for 'telephone', as there is no real business connected with the medium, but for 'newspaper', this line of reasoning is not applicable. Here, we might argue that the profession of making news(papers) is denoted by other nouns, such as 'journalist' and 'publisher', so that

⁵Here, and in other places (always indicated by a footnote), the sample was found too big to fully categorize by hand, and it might well be that an interesting category has been missed.

the bare PP 'in newspaper' is overridden by these nouns. For 'radio', 'film' and 'television', such names seem to be less familiar.

With regard to (2.26), [Paenen, 2009] states that in Dutch, 'in' licenses noun of the CLOTHING-type (e.g. 'in rok' - 'in robe', 'in pak' - 'in suit'). In the Brown corpus, we only find 'in uniform' and 'in shirt'. We could jump to the conclusion that this class is less productive in English. However, when we use Google to find extra examples, we see that some translations of the nouns that Paenen found in Dutch are also found in English (e.g. 'in suit' is perfectly acceptable but does not appear in the Brown corpus). Paenen notes that the 'in CLOTHING' does not seem totally productive, even in Dutch: 'in jurk' ('in dress') is at least questionable.

In this case, the real test would be to compare the results within a multilingual corpus, which should ideally include a session on current fashion. However, since such corpora do not yet exist, such a search is outside the scope of these thesis. From the examples I did come across, I observe that 'in' seems to only occur determinerless with nouns of the CLOTHING-category when the expression refers to a particular dress code ("Everybody has to come in suit to the party"). This change in semantics is typical for bare PPs, as we have noticed several times before.

Furthermore, one should note the difference between the categories (2.25) and (2.29): while 'in' licenses nouns which refer to a long period of time, 'by' only refers to short periods or sudden changes. These differences are therefore due to the semantics of the preposition: 'in' demands the noun to be referring to an interval, so that it can refer to a period within this interval, while 'by' wants the noun to be referring to a single occurrence. A similar account might be made as to why the nouns in (2.30) might appear bare after 'in': these denote several 'intervals' in life.⁶

Generalizing this observation, we would expect that with prepositions like 'at', 'before' and 'past', we would find similar TIME nouns as was the case with 'by', since all these prepositions are non-inclusive. And indeed, we find that only nouns of the SHORT PERIOD OF TIME-category combine with the non-inclusive prepositions: see Table 2.5 and the accompanying text there.

There are, however, some cases in which 'by' does appear with a longer period of time ('by year', 'by week', 'by month'). However, when we look at the context in which these PPs appear throughout the corpus, we see that 'by N' appears in a construction 'N by N'. There are actually quite a few examples of this sort: not only 'day by day' and 'year by year' exist, but also 'side by side', 'step by step' or even 'dance by dance'. [Jackendoff, 2008] has written a comprehensive article on this N-P-N construction. It is however questionable whether this article sheds any new light on the semantics of the preposition with respect to the more frequent P-N constructions: the N-P-N constructions seem to be in a league of their own.

2.2.4 Conclusion

For English, it is clear that there is a myriad of prepositions licensing bare nouns. First of all, there is 'per', that can take any bare noun as its complement, which then acquires a UNIT reading (but a specific semantics seemed hard to find). Next, the preposition 'under' in English licenses a specific type of nouns: nominalizations of transitive verbs, yielding a reading which is likewise to the passive. Concerning 'in' and 'by', we have seen some interesting productive patterns: 'in' opened a lexical slot for nouns of both the CLOTHING and the MEDIA AS BUSINESS-category, and 'by' opened slots for any MEANS OF TRANSPORTATION and any COMMUNICATION MEDIA. Here, we also saw that the nouns of the TIME-category they occur with are complementary: while 'in' combines with long periods of time, 'by' only takes short periods of time as its complement. The next question is: will we find that 'with' and 'without' also allow for bare nominals as their complement?

⁶Again, I must note that most of the nouns in these categories are not strictly countable. Still, I think the observation might be of interest, even though [Haspelmath, 1997] already posed somewhat similar conclusions.

2.3 \exists -based bare PPs

When one has digested the data in the graph and table above, one thing might have struck the attention. While 'without' is ranked as one of the prepositions that is eager to appear with a bare noun, the antonym 'with' is only to be found halfway the ranking. This is odd: we would expect these two prepositions to be on the same level, as there seems to be little difference in semantics on the surface between the two prepositions. If we scrutinize the occurrences in more detail, the plot only thickens. When we apply the method of Kiss (introduced in the beginning of this chapter) for determining whether or not a noun is countable (while not looking at the whole sentence), we observe that for 'without', slightly more than 37% of the nouns following this preposition are countable (applying the same method to Dutch yields a percentage of 45% bare singular count nouns after the preposition). For 'with', this measure is close to 18% (in Dutch: 23%).

If we however look more closely at the corpus data, and do take a look at the sentences as a whole, we see that it is with 'with' that we find a lot of pollution in the data. For almost all of the bare nouns appearing after 'with' we see that these nouns are either used as an adjective (in 2.31) or in bare coordination (in 2.32). Other examples involve crimes like 'murder' and 'riot', which are only questionably singular count nouns in the uses in (2.33) and (2.34) below.⁷

(2.31) They had been fed a hunting breakfast, so called because a kedgerree, the dish identified **with fox hunting**, was on the bill.

(2.32) I want the fish served whole, **with head and tail**.

(2.33) I would have been negligent and a goddam lousy cop to boot, if I'd sat around this station all night when somebody got away **with murder** in my district.

(2.34) When Sir Edward Greville enclosed the town commons on the Bancroft, Quiney and others leveled his hedges on January 21, 1600, and were charged **with riot** by Sir Edward.

In the Brown Corpus, we only find only one example of 'with' followed by a (real) bare singular count noun. I have copied the example in (2.35) below. Another idiomatic use of 'with' (not found in the corpus) would be 'with child', meaning pregnant. One might conclude that 'with' is highly unlikely to appear with a bare singular count noun as its complement.

(2.35) It encompasses in its expanse areas where the natural beauty encourages a vacation of quiet contemplation, on the one hand, to places where entertainment and spectacles of all sorts have been provided for the tourist **with camera**.

For 'without', on the other hand, we see that there are more examples to be found:

(2.36) without exception, analysis, comment, question, accusation, argument

Still, it seems that these nouns also form a restricted lexical class. It seems that one cannot simply plug in any bare singular count noun after 'without'. If that would be the case, I would expect the percentage of bare singular count nouns following 'without' to be at a similar level as that of Dutch, in which bare nouns seem to appear freely after 'without', and which led [Le Bruyn et al., 2009] to pose no restriction on the lexical class of the noun in the semantics for \exists -based bare PPs (see chapter 1). Again, only one example seems to prove me wrong in this case:

(2.37) – that should a minister in Boston trust himself to his heart, should he "speak **without book**, and consequently break some law of speech, or be hurried into some daring hyperbole, he should find little mercy".

⁷This is clearly one of the drawbacks of our rough acquisition of the data. Only late in the analysis we find that the extracted bare PP is actually not a bare PP. The method used here therefore requires a lot more manual labor than one would have initially expected.

What *is* clear, on the other hand, is that ‘with’ is not suitable in any of the contexts where we find ‘without’. To prove this simple fact, I have added some examples from the corpus.

(2.38) Charlie had accepted the diagnosis *with/without comment.

(2.39) The stepmother, almost *with/without exception, has been presented as a cruel ogress.

What might be an explanation for this deviation? It might be that negative contexts are more likely to contain bare nouns, as [Baldwin et al., 2006] seem to think. After they report that for bare mass nouns, the count for ‘without’ was roughly double that of ‘with’, they conclude that “cross-cutting semantic features such as [...] negative polarity play[s] a role in the semantic regularity of [bare PPs]”. In two footnotes of [Longobardi, 1994] (number 6 and 13), a similar argument is put forward, though Longobardi claims that this Italian “ability of negation to license an empty D” is only true for “certain lexical choices” (e.g. ‘non ha proferito verbo’ (‘he didn’t utter word’) is grammatical, while ‘non ha dipinto quadro’ (‘she didn’t paint picture’) is out.).

Another (new) idea for this deviation between ‘with’ and ‘without’ is more on the semantic level. It might be that ‘without’ is more likely to be followed by bare singular count nouns, because the semantics for ‘without + N_{sg} ’ is likewise to that of ‘without + N_{pl} ’. For e.g. ‘without exception(s)’, both the singular and the plural form state that there are no exceptions. For ‘with’, this relationship does not hold: with a plural after ‘with’, more instances of the noun are being referred to.

However, all of these claims are done on a minimal amount of data, and do not in any way explain the behavior found. Furthermore, it is striking that in both Dutch and French, the translations of ‘with’ and ‘without’ *do* seem to appear freely with bare nouns. We would also like to have an explanation for that fact. In the next chapter, I will review examples of Dutch uses of ‘met’ and ‘zonder’, the direct translations of ‘with’ and ‘without’. In the chapter on multilingual analysis per preposition, I will try to further compare the uses of ‘without’/‘zonder’ in English and Dutch. In this manner, we might be able to say more on this particular matter than we are able to do at this stage of the thesis.

2.4 N-based bare PPs

In the last section, we have searched for bare PPs which were mostly dependent on the preposition in use. We have seen, however, that there are also nouns that occur bare with more than one preposition. This suggests that it is the *noun* that plays a role in its own licensing. This category of bare PPs is consequently conveniently named ‘N-based bare PPs’.

In English, we find that there are actually a lot of nouns of this type. In the Brown Corpus, there are over 1500 nouns that appear with more than one preposition. At first sight, these nouns do not seem to restrict themselves to the ‘place’ and ‘time’ (respectively ‘places’ and ‘interruptions’ in Stvan’s vocabulary) categories of Dutch and French, reported by [Le Bruyn et al., 2009]. After extracting those nouns which appear with more than one preposition in the Brown Corpus, I have tried to divide these nouns manually in several subclasses. In the following sections, I will present some of these subclasses, and carefully go over the data provided, because again, there is a possibility that we end up with some pollution in the data.

2.4.1 Locations

The first subclass from the familiar locations of [Stvan, 1998]. We see that here, especially the preposition ‘in’ is prevalent. In the cases in which ‘in + N’ would yield an ungrammatical utterance (as is the case with ‘in stage’ and ‘in earth’), the alternative is ‘on’. However, as is clear from the examples, the nouns here are by no means limited to any preposition: ‘school’ can appear bare with as much as 13 prepositions. Note that the fact that ‘school’ appears after a preposition like ‘during’ is a strong argument for a reading in which ‘school’ indeed denotes an activity, something that takes some time, as was already claimed earlier in this thesis.

Table 2.2: Nouns of the LOCATION category with (1) the number of occurrences of the noun in bare PPs, (2) the number of distinct prepositions the noun occurs with and (3) a specification of both columns: which preposition-noun combinations do we find and in what number? (This table is shortened for reasons of space, thus providing only the nouns which appear with the highest amount of distinct prepositions.)

noun	#	# d.p.	preposition (# of occurrences with noun)
school	76	13	about (1), after (4), against (1), at (4), between (1), during (1), for (3), from (5), in (22), of (16), through (1), to (16), with (1)
work	75	11	about (1), at (14), by (2), for (10), from (2), in (1), of (24), off (1), on (2), to (17), toward (1)
town	62	10	after (1), around (1), for (3), in (23), into (5), of (13), through (2), to (13), towards (1)
college	37	9	after (1), among (1), at (3), for (2), in (13), of (5), through (3), to (8), with (1)
stage	22	9	across (1), among (1), at (1), by (2), for (2), from (2), of (6), on (6), to (1)
earth	41	8	between (1), from (3), into (1), of (9), on (17), over (1), to (5), with (4)
court	20	8	at (4), from (1), in (8), into (1), of (3), to (1), under (1), without (1)
home	52	7	at (31), during (1), for (2), from (11), in (3), of (3), with (1)
prison	19	7	between (1), by (1), from (2), in (11), of (1), on (1), to (2)
church	32	6	at (3), for (6), from (2), in (6), of (5), to (10)
space	29	6	by (1), in (16), into (3), of (7), than (1), through (1)
bed	47	5	for (1), in (18), into (5), of (8), to (15)

2.4.2 Media

For nouns of the MEDIUM category, we see that again, there are some nouns which can appear with a multitude of prepositions. I have copied the nouns in Table 2.3 below. Note that 'TV' and 'television' of course could here have been merged together, but to improve the clarity of this corpus experiment (should anyone want to duplicate it), I have chosen not to.

Especially 'radio' might be of interest here. We already saw 'radio' as object of the prepositions 'by' and 'in', but it seems that 'radio' can appear with a broader range of prepositions. On closer scrutiny, we however find that the amount of prepositions it forms a good bond with is actually less than is given in the frequency table in 2.3 (mostly due to composites like 'radio waves' and the like). Still, we get some interesting forms in the examples below:

- (2.40) For 10 years Sponsor has issued an annual survey of the size and characteristics of the Negro market and of successful techniques for reaching this market **through radio**.
- (2.41) The survival of the family will depend largely on information received **by radio**.
- (2.42) In recent days there have been extensive lamentations over the absence of original drama on television, but not for years have many regretted the passing of new plays **on radio**.
- (2.43) KARL provides experience for students who wish to pursue careers **in radio**.

It seems that in each of the sentences above, another property of 'radio' is being put to use. In (2.40) and (2.41), 'radio' is seen as an object of communication, though in the latter, the radio still is seen as a physical object. In (2.42), 'radio' seems to be more the total of programs that are being broadcast by the radio. And in (2.43) (which we came across earlier when we analyzed the preposition 'in' for possible categories of P-based bare PPs), we see that 'radio' is seen as

Table 2.3: Nouns of the MEDIA category with (1) the number of occurrences of the noun in bare PPs, (2) the number of distinct prepositions the noun occurs with and (3) a specification of both columns: which preposition-noun combinations do we find and in what number?

noun	#	# d.p.	preposition (# of occurrences with noun)
radio	18	9	at (1), by (3), for (1), in (3), of (4), on (3), through (1), to (1), with (1)
film	19	7	after (1), against (2), in (2), of (10), on (1), to (2), upon (1)
television	17	7	after (1), in (1), into (1), of (6), on (6), to (1), with (1)
newspaper	7	6	for (1), from (1), in (1), of (2), outside (1), with (1)
paper	31	5	in (2), of (17), on (9), to (1), with (2)
TV	11	4	before (1), for (1), of (4), on (5)
tape	6	3	of (2), on (2), with (2)
mail	4	3	by (1), of (2), with (1)
telephone	7	2	by (5), of (2)
phone	3	2	by (2), of (1)

a business. I will argue from these examples that for N-based bare PPs, a well-known idea of [Pustejovsky, 1991] might be of help to find a proper semantics. Pustejovsky argues that nominals have a **qualia structure**. The meaning of a noun consists not just of its denotation in the world, but a word has four roles:

- the relation between it and its constituent parts (Constitutive Role)
- its physical characteristics (Formal Role)
- its purpose and function (Telic Role)
- whatever it brings about (Agentive Role)

An enlightening example of where this qualia structure could come in use is the difference between the following two examples:

(2.44) John baked the potato.

(2.45) John baked the cake.

In the former, we see that the potato undergoes a transformation via the process of baking. While indeed, in the latter example, the cake also undergoes a transformation, the sentence also presumes that John is the one who *created* the cake. We could now do two things: either have two different entries for the verb 'to bake' (one a state-transforming sense, the other the creating sense), or, as Pustejovsky argues, put some of the semantic weight on the noun phrase by using qualia structures. This latter approach has a certain elegance, since the verb is not regarded as being polysemous and having several entries in our mental dictionary.

For a noun like 'radio', we would then acquire these four roles:

- Constitutive Role: *medium(x)*
- Formal Role: *object(x)*
- Telic Role: *communicate(y, x)*
- Agentive Role: *listen(P, x), broadcast(P, x)*

With the help of the examples above, we could now make the suggestion that the dropping of the determiner signals that either the telic or agentive role of the noun should be used. With a full PP ('on/near/behind the radio'), we find that the physical characteristics of 'radio' (and thus its constitutive and formal role) are being referred to. In this way, the semantic weight would be on the noun, and not on the preposition, just as we wanted from the beginning.

2.4.3 Parts of Body

In the PARTS OF BODY category, we also find nouns that appear with several distinct prepositions. I have here added some nouns that are neither countable (e.g. 'mind') nor really a part of the body (e.g. 'sight' and 'touch', which are, of course, senses).

Table 2.4: Nouns of the PARTS OF BODY category with (1) the number of occurrences of the noun in bare PPs, (2) the number of distinct prepositions the noun occurs with and (3) a specification of both columns: which preposition-noun combinations do we find and in what number?.

noun	#	# d.p.	preposition (# of occurrences with noun)
hand	62	8	at (14), by (5), from (1), in (22), of (5), on (13), to (1), with (1)
view	62	7	beyond (1), in (19), into (6), of (33), on (1), to (1), within (1)
heart	12	7	at (3), by (1), in (1), of (3), to (1), with (2)
head	32	6	from (3), of (5), over (1), per (20), to (1), with (2)
sight	27	6	from (3), in (9), into (2), of (11), on (1), within (1)
breath	11	4	for (7), of (2), per (1), to (1)
eye	4	4	at (1), by (1), of (1), to (1)
mind	65	3	in (34), of (23), to (8)
touch	24	3	by (2), in (11), of (11)
foot	6	3	of (2), on (3), to (1)
mouth	4	3	in (2), of (1), on (1)
pupil	4	3	against (1), of (2), with (1)
body	12	2	in (4), of (8)
arm	2	2	for (1), in (1)
feeling	2	2	of (1), with (1)
flavor	2	2	in (1), of (1)

However, when we take a closer look at the examples, for example that of 'hand' and 'heart', we see that these are sometimes idiomatic, while on the other hand sometimes being more compositional. A primary example form (2.50) and (2.51): in the former sentence, 'in hand' means 'under control', in the latter 'in hand' is taken far more literally. From the examples, it does not seem that Pustejovsky's qualia structure might be of help here.

(2.46) About 500 employees of the firm will be **on hand** to witness bestowal of the honor upon Mr. Sorrentino.

(2.47) Since the work is done **by hand**, the only limitation, it is said, "is that of human conception".

(2.48) In the light of the facts **at hand**, however, New York Central intends to pursue the objective of helping to create a healthy two-system eastern railroad structure in the public interest.

(2.49) When things got a little **out of hand**, they very rapidly got a lot **out of hand** – it seemed to be a general rule.

(2.50) At that time, the people at the bank said they felt that they had the situation **in hand**.

(2.51) With the Fund **in hand**, the debt on the boilers had been paid.

Nonetheless, it might be that the dropping of the determiner signals an idiomatic meaning. One could imagine a grammar that says: if a determiner were present, then we should have a locative semantics for the preposition-noun-combination, otherwise, the semantics is fully idiomatic. However, there are two major problems with this approach. First of all, for bare PPs with nouns like 'school', 'prison' and 'church', we also find bare forms, but here the semantics is far from idiomatic, it follows a quite predictable pattern, as has been argued some times before in this thesis. We would then have two similar forms (prepositions followed by nouns missing a determiner) with distinct semantics (one regular, one idiomatic), and this is certainly not optimal from a hearer's point of view.

Secondly, and possibly more importantly, idiomatic forms are (at least in Dutch) also found with full PPs, as one of my reviewers pointed out. For example: having someone 'op het oog' ('on the eye') does not mean that there is a person that is physically on one's eye, but that one knows someone who can do the job. And following someone 'op de voet' ('on the foot') again does not mean that one is on another's foot, neither that one is following another on foot, but rather that one is close behind the other. These examples form strong arguments against the idea that the dropping of the determiner signals an idiomatic meaning. The conclusion from this section should then be that the PARTS OF BODY-category is not a legitimate candidate for N-based bare PPs.

2.4.4 Times

In this class, we find the timely interruptions of [Stvan, 1998] (like 'breakfast' and 'lunch'), but also times of day and longer periods. While this latter category of nouns are not strictly within the realm of bare nominals, they do form an interesting couple: we see that shorter periods of time appear with other prepositions than the longer periods of time, as is also specified in Table 2.5. As is clear from this table, longer periods (like the seasons, 'year', 'darkness' and 'daylight') mostly combine with prepositions like 'in', 'during' and 'by', while shorter periods can appear bare with prepositions like 'at', 'after', 'before' and 'by'.

2.4.5 Conclusion

The conclusion from this section on N-based bare PPs in English should be that indeed, Stvan was right: there seem to be only two classes of nouns that can appear bare with several prepositions. First of all, we find the locations, where we find that the dropping of the article yields an enriched semantics. And secondly, there are the times of day and interruptions, which most of the time do not have a non-bare (full) counterpart. However, using our corpus research we were able to establish that some nouns only prefer a limited range of prepositions.

For other categories of nouns, there are indeed some nouns that can appear with several prepositions, but here it seems that we are mostly dealing with idiomatic expression (for example within the PARTS OF BODY-category). For the semantics of some N-based bare PPs (like 'radio'), we embraced the Pustejovsky's theory on the Generative Lexicon. With some examples, I showed that such a theory could account for the various meanings the noun seems to acquire.

2.5 Conclusion

The corpus research into English bare PPs yielded some interesting results. First of all, concerning P-based bare PPs, we found that English allowed for constructions, some of them already to be found in the literature (e.g. 'by + MEANS OF TRANSPORTATION was already reported by [Baldwin et al., 2006]), other constructions (like 'under + ACTION' and 'in + MEDIUM') have been recognized for the first time in this thesis. Further cross-lingual research should find out whether we also find these constructions in other languages.

Table 2.5: Nouns of the TIME category with (1) the number of occurrences of the noun in bare PPs, (2) the number of distinct prepositions the noun occurs with and (3) a specification of both columns: which preposition-noun combinations do we find and in what number? In the last column, a specification of whether the time is long or short is added.

noun	#	# d.p.	preposition (# of occurrences with noun)	long/short?
summer	21	11	about (1), before (1), by (1), during (1), for (1), from (1), in (8), into (1), of (4), until (1), with (1)	long
midnight	17	8	after (1), at (7), before (3), between (1), by (1), from (1), past (2), to (1)	short
dinner	29	7	about (1), after (2), at (9), before (2), for (6), to (7), with (2)	short short
night	59	6	after (2), at (46), before (1), by (2), for (3), of (5)	short
day	41	6	after (3), by (4), from (4), of (5), per (21), to (4)	long
noon	16	6	after (2), around (1), at (8), before (1), by (2), to (2)	short
dawn	15	6	at (5), before (6), by (1), from (1), of (1), since (1)	short
spring	13	6	between (1), by (1), for (1), in (3), of (6), on (1)	long
breakfast	13	5	at (2), before (4), for (5), of (1), to (1)	short
week	9	5	after (1), by (1), from (2), per (3), to (2)	long
sunrise	7	5	after (3), around (1), before (1), from (1), of (1)	short
sunset	7	5	after (2), around (1), before (2), near (1), to (1)	short
darkness	7	5	in (2), into (2), of (1), on (1), with (1)	long
year	19	4	after (1), by (1), of (3), per (14)	long
luncheon	5	4	after (1), at (2), before (1), by (1)	short
nightfall	4	4	after (1), at (1), before (1), by (1)	short
winter	14	3	for (2), in (7), of (5)	long
month	9	3	after (1), by (1), per (7)	long
supper	8	3	after (3), at (1), for (4)	short
lunch	6	3	after (2), for (3), to (1)	short
sunlight	6	3	in (1), into (1), of (4)	long
dusk	5	3	after (1), at (3), to (1)	short
daylight	5	3	before (1), during (3), in (1)	long
autumn	5	3	for (1), in (1), of (3)	long
sundown	4	3	after (1), before (2), by (1)	short
morning	3	3	before (1), from (1), of (1)	short
afternoon	3	3	by (1), of (1), with (1)	short
midday	2	2	at (1), of (1)	short
midsummer	2	2	in (1), to (1)	long
springtime	2	2	by (1), in (1)	long

Concerning N-based bare PPs, we found that apart from the categories already mentioned by [Stvan, 1998] (viz. locations, timely interruptions and in a lesser amount media), we had a hard time finding new categories of nouns that can appear bare after a multitude of prepositions. A potential new category would be the PARTS OF BODY-category, but on closer scrutiny we found that these nouns are in fact quite limited with regard to the prepositions they appear with, and furthermore often form more idiomatic expressions. Research into the two already established categories did however yield some interesting results: concerning nouns in the TIMES-category, we found a nice distinction between shorter and longer periods in prepositions preceding the noun.

With regard to \exists -based bare PPs, we found that English does not seem to allow these for 'with'. We were not able to find any convincing examples of bare nouns after this preposition. For its logical counterpart, 'without', we did find some examples. However, it is questionable whether 'without' is as free as e.g. Dutch in the allowance of bare nominals as its complement. Further monolingual and multilingual analysis should help shedding some light on these issues.

With these results, we are able to make a typology for the three types of bare PPs in English. I have done so in Table 2.6 below. An important design choice here was that I have put the nouns of the MEDIA-category in separate P-based-classes, while for locations, I made a N-based-class. I believe this categorization is warranted, because we saw that with nouns in the MEDIA-class, there seem to be quite some differences in the prepositions they combine with.

It would be very interesting to fill and/or extend this table with results from monolingual analysis of other languages. After that, we are able to find interesting categories of bare PPs for multilingual research. In the next chapter of this thesis, we will therefore analyze the occurrences of Dutch bare PPs.

Table 2.6: A typology for bare PPs in English. In the first column, we find the three types of bare PPs, in the second column, the categories of nouns (or preposition for \exists -based bare PPs), and the last column gives the preposition English uses for this form (for P-based bare PPs) or the number of occurrences (for N-based and \exists -based bare PPs). These latter 'numbers' (many, some and zero) are only indicative, and should not be interpreted with any real number.

P-based bare PPs	Category of nouns	English preposition
	N \Rightarrow UNIT	per
	MODES OF TRANSPORT	by
	COMMUNICATION MEDIA	by
	MEDIA AS BUSINESS	in
	CLOTHING	in
	ACTION	under
N-based bare PPs	Category of nouns	# in English
	LOCATION	many
	INTERRUPTION	many
\exists-based bare PPs	Preposition	# in English
	<i>with</i>	zero
	<i>without</i>	some

Chapter 3

The Dutch Data

3.1 Introduction

In this chapter, I will present the data for Dutch in a similar fashion to the previous chapter on English. For the Dutch analysis, I will include earlier research of both [Le Bruyn et al., 2009] (concerning P-based bare PPs) and [Paenen, 2009] (concerning N-based bare PPs). However, there are at least two reasons to include a chapter on Dutch in this thesis. First of all, there is reason for a review of at least some of the data provided by these two authors. For example, Paenen 'only' researched those categories of nouns (locations and interruptions) that also appeared in [Stvan, 2007]. It might be that a research along the lines of the previous chapter yields some categories of nouns yet unaccounted for. And Le Bruyn et al. only provided analysis of three prepositions, so it might be that further analysis yields more productive P-based bare PPs. The second reason for taking a look at Dutch is equally important: we are in need of this data to get a clearer look on interesting bare PPs for a multilingual analysis. Without first performing a monolingual analysis, there is some risk that interesting forms are being overlooked.

3.1.1 Method

Once again, I used the method of [Le Bruyn et al., 2009] to acquire the results for Dutch. This time, completely similar to Le Bruyn et al., the Eindhoven Corpus ([van Grootheest, 1989]) is used to find bare PPs. This corpus consists of both written texts and transcribed conversations. The corpus contains around 768.000 tokens. Importantly for our purposes, it is tagged for part of speech.

Again, we search for prepositions followed by a bare noun. Within the Eindhoven corpus, this means that the search query (a so-called *regular expression*) should be as follows:

(3.1) '[a-z]*_600 [a-z]*_000'

So, we search for any ('[a-z]*' is the regular expression for zero or more letters between a and z, i.e. words) preposition (marked by the tag '_600' in the Eindhoven Corpus) followed by any (again '[a-z]*') singular noun ('_000'). Each hit is then stored in a database table with two columns, preposition and noun. Using Structured Query Language (SQL) we are able to group the data in any way we like.

3.1.2 Drawbacks to the Method

As the POS-tagging for the Eindhoven Corpus is quite similar to that of the Brown Corpus, the method described in the above section has similar drawbacks to that of the previous chapter. To refresh our memory: first of all, we should watch our back for constructions that share properties

with bare PPs, but are far out of the scope of the current research, as their role in both syntax and semantics seems unrelated to bare PPs (e.g. N-P-N-constructions).

A second hurdle is formed by the mass-count distinction. Since the POS-tagging does not make a distinction between the two, it is likely that the statistics provided are somewhat polluted by mass nouns.

3.2 P-based bare PPs

As stated already a few times in this thesis, P-based bare PPs are essentially prepositions which open a slot for a noun of a specific lexical category to appear bare. In the section on English P-based bare PPs, we adopted the method of [Le Bruyn et al., 2009] to find this kind of bare PPs. In this section, we will again employ this method to find Dutch P-based bare PPs (under the motto: never change a winning method). To refresh our minds: Le Bruyn et al. counted per preposition (1) its total number of occurrences (2) its number of occurrences followed directly by a singular noun. Dividing the latter by the former then gives us the relative frequency of bare nouns following the preposition. We then rank every preposition according to this measure, and find that the prepositions on top of the table are suitable candidates for being a P-based bare PP.

In Table 3.1, we find the results of this procedure. A graphical representation of this table can be found in Figure 3.1 below.

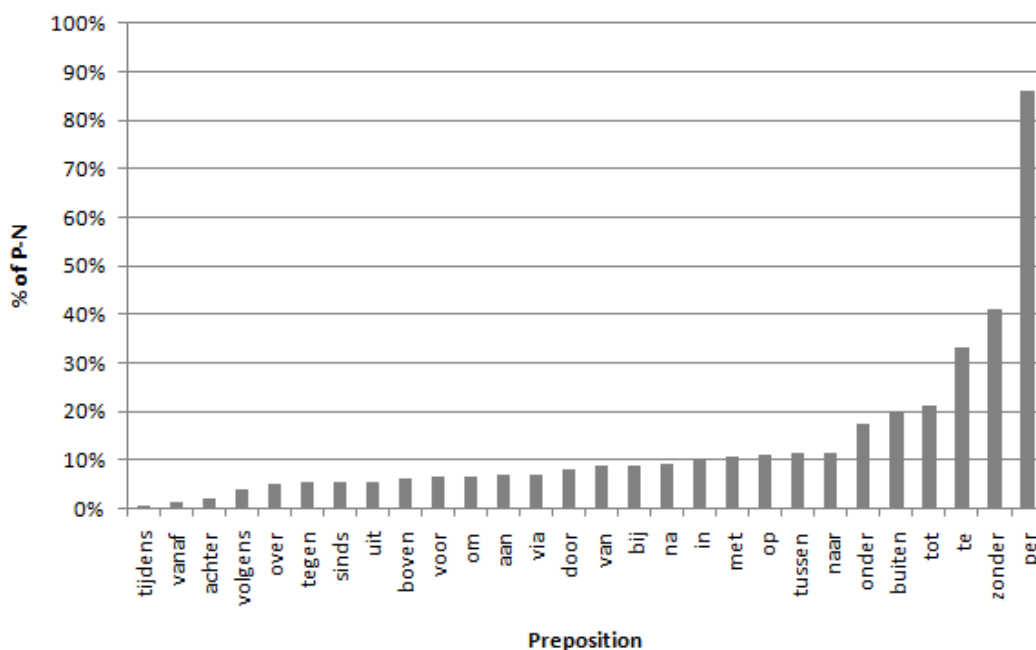


Figure 3.1: Relative frequency of bare nouns per preposition in Dutch (for translations of the prepositions, see Table 3.1).

It cannot be noted enough that this is quite a rough procedure to extract the data from the corpus. Since one only searches for $[P N_{singular}]$ patterns, one might also include all kind of unwanted patterns into the statistics. The number and percentages are therefore only indicative: one should not try to draw any statistically significant conclusion from the data in Table 3.1 and the subsequent graph. On the other hand, this preprocessing of the data *does* help to direct us towards interesting bare PPs. In the next sections, I will therefore take a closer look at some of the prepositions in the table.

Table 3.1: Frequency data for English. From left to right: preposition, number of occurrences, number of bare nouns, percentage of bare nouns. In the penultimate column, we find the noun the preposition occurs most common (m.c.) with. For reasons of space, less common prepositions have been omitted from this table. Note that for four prepositions ('om', 'onder', 'over' and 'uit') the results are somewhat different from [Le Bruyn et al., 2009], due to a stricter search query on my behalf. Furthermore, I have added the preposition 'te' to the table.

P	#	# BN	% BN	m.c. BN	#
per (per)	273	317	86,12%	jaar (year)	56
zonder (without)	145	352	41,19%	twijfel (doubt)	7
te (to)	140	421	33,25%	werk (work)	20
tot (until)	491	2339	20,99%	stand (position)	68
buiten (outside)	30	151	19,87%	beschouwing (consideration)	8
onder (under)	139	795	17,48%	leiding (leadership)	30
naar (to)	308	2665	11,56%	huis (home)	79
tussen (between)	75	668	11,23%	kerk (church)	3
op (on)	757	6861	11,03%	grond (ground)	64
met (with)	728	6761	10,77%	betrekking (regard)	77
in (in)	1512	15500	9,75%	staat (state)	95
na (after)	73	790	9,24%	afloop (end)	14
bij (at)	258	2911	8,86%	voorbeeld (example)	36
van (of)	1885	21473	8,78%	mening (opinion)	43
door (through)	230	2816	8,17%	middel (means)	43
via (via)	12	174	6,90%	kleuring (coloring)	1
aan (on)	263	3878	6,78%	boord (board)	24
om (around)	49	741	6,61%	hulp (help)	8
voor (before)	356	5470	6,51%	stuk (piece)	9
boven (above)	10	165	6,06%	water (water)	6
uit (out)	99	1805	5,48%	zee (sea)	9
sinds (since)	7	132	5,30%	jaar (year)	3
tegen (against)	49	926	5,29%	betaling (payment)	5
over (over)	97	1939	5,00%	straat (street)	3
volgens (according)	12	307	3,91%	opgave (task)	2
achter (behind)	5	269	1,86%	winnaar (winner)	1
vanaf (from)	1	74	1,35%	woensdagmorgen (W. morning)	1
tijdens (during)	1	203	0,49%	vakantie (vacation)	1

The method for this 'closer look' was already explained in the previous chapter, but is copied (in abbreviated form) for the sake of clarity:

1. Selecting the nouns (and their count) occurring with the preposition under study.
2. Disregarding all clear idioms from the selection.
3. Disregarding all non-countable (mass) nouns from the selection.
4. Categorizing the remaining bare singular count nouns by hand.

In the now following sections, I will report the results per preposition. I will start off with the highest preposition in the table when ordered on percentage of bare nouns, 'per', and then work down the table, skipping some of the prepositions for which I expect a large amount of either idioms (e.g. with 'tot' and 'buiten') or mass nouns following the preposition (e.g. with 'volgens').

3.2.1 *Per* as Polysemous Preposition

'Per' in Dutch quite clearly has far more functions than its English counterpart. While it again licenses every singular count noun, yielding a 'unit' reading (see for example (3.2) and (3.3)), under the right circumstances, 'per' can yield some quite different readings. For Dutch, for example [Le Bruyn et al., 2009] show that nouns of the category MEANS OF TRANSPORTATION can follow 'per'. Most of the time, these forms do not yielding a 'unit' reading, but rather a reading where English speakers would apply the preposition 'by'. Some examples of this behavior (taken from the Eindhoven Corpus) are provided below in (3.4) and (3.5).

(3.2) *Per regio treden er aanzienlijke verschillen op.*
Per region come there substantial differences at.
"There are substantial differences per region."

(3.3) *Tussen Amsterdam en Schiphol zal de lijn ruim twaalfmiljoen reizigers per jaar vervoeren.*
Between Amsterdam and Schiphol will the line more than twelve million
travelers per year transport.
"Between Amsterdam and Schiphol, the line will transport twelve million travelers a year."

(3.4) *Vier jonge Rotterdammers willen deze zomer per auto naar Japan.*
Four young inhabitants of Rotterdam want this summer per car to Japan.
"Four young inhabitants of Rotterdam want to go to Japan this summer by car."

(3.5) *Ik ging niet per tram, maar besloot op mijn gemak te gaan lopen.*
I went not per tram, but decided at my ease to go walk.
"I didn't go by tram, but decided to take an easy walk."

Not only means of transportation, but also members of the MEDIUM-class can appear after the Dutch 'per'. For example, one can tell one another that a message will be send 'per e-mail' ('by e-mail') or it will be conveyed 'per SMS' ('by SMS'). Some other examples are provided in (3.6) and (3.7) below. However, it seems that Dutch speakers often prefer to use the construction 'via (+ DET) + MEDIUM' ('via de telefoon' has almost three times the amount of hits of 'per telefoon' on Google).

(3.6) *Waarom zou je, als je dat per telefoon kunt doen, al die moeite doen om een brief te schrijven?*
Why would you, if you that per telephone could do, all that effort
do to a letter to write?
"Why would take all that effort to write a letter, if you could handle all this by telephone?"

- (3.7) *De andere organisaties zijn gisteren per brief van de schenking op de hoogte gesteld.*
The other organisations are yesterday per letter of the

“The other organisations have yesterday been by letter.”

Another distinct and widely used form of ‘per’ in Dutch is that of ‘per ongeluk’ (‘by accident’ or ‘accidentally’). However, it is questionable whether there are more nouns of this category, in other words, whether we have stumbled upon a productive category of nouns. In Dutch, I am only aware of ‘per abuis’ as a likewise construction with a similar meaning. In English we find these forms translated ‘by mishap’ and ‘by mistake’ (this latter form is strictly translated into Dutch with ‘per vergissing’, which seems grammatical, but seems only to be widely used in Flemish).

Yet another category: ‘per’ might also be followed by some very near time in the future: ‘per direct’ (‘immediately’), ‘per ommekeer’ and ‘per kerende’ (both most easily translated with ‘by return’) are the primary examples. The last example is most of the time followed by a MEDIUM. Nowadays, this class seems productive: while in the Eindhoven Corpus, only ‘per kerende post’ (‘by return mail’) was found, a Google search now returns also ‘per kerende telefoon’, ‘per kerende pakketpost’ or even ‘per kerende SMS’ and ‘per kerende e-mail’.¹ With these last examples, it seems like not ‘per’, but ‘per kerende’ is the preposition which opens up the slot for the bare noun, as some of my reviewers remarked.

Lastly we find that ‘per’ in Dutch plays a role in some more idiomatic expressions, borrowed from Latin: ‘per saldo’, which would mean something like ‘ultimately’ and ‘per se’ which means ‘necessarily’. In English, ‘per se’ seems to have a bit different connotation (‘without consideration of extraneous factors’, as in “The law makes drunk driving illegal per se.”). However, as both forms are not productive in any way, these forms are not really of interest in this thesis.

To conclude: ‘per’ in Dutch is more polysemous than in English, as the prepositions also licenses the bare occurrence of modes of transport and means of communication. English here uses ‘by’. Concerning the UNIT reading that other nouns acquire after ‘per’, we have not yet got any clearer view: maybe a multilingual analysis will help to acquire a more insightful analysis on that matter.

3.2.2 *Te: Bare*

The preposition ‘te’ (most directly translated with ‘to’) is also high in the list of P-based bare PPs. The preposition ‘te’ is comparable to ‘per’ in its impossibility to appear followed by an article, which might explain its high rank in Table 3.1 above. On closer scrutiny, the nouns following ‘te’ seem, unlike ‘per’, to belong to a closed category, and the bare PP as a whole must be in the presence of a light verb. Consequently, it seems probable that ‘te’ actually is used to link the noun and the verb.

Examples of the nouns following ‘te’ are to be found in (3.8) and (3.10) below (translations in (3.9) and (3.11)). In the former examples, we find ‘te’ followed by nouns of the MEANS OF TRANSPORTATION-category. However, these two nouns are a deviation of the norm, as ‘per’ is used for all other modes of transport. In English, we also see that a different preposition (namely ‘on’) is used for these nouns. The latter category indeed contains bare PPs which only appear with a specific or limited amount of verbs. An example is ‘te woord staan’ (literally ‘to word stand’, ‘to speak to’): another verb would here render the bare PP ungrammatical.

- (3.8) *te voet, te paard.*

- (3.9) *on foot, on horseback.*

¹An example like ‘per kerende trein/bus’ (‘by return train/bus’) also is grammatical (but not ‘per kerende fiets’ (‘by return bicycle’)) so it might be that we should extend the class of nouns which can appear after this construction also to some means of transportation.

(3.10) te water (gaan), te werk (gaan), te lijf (gaan), te hulp (komen/schieten), te koop (staan), te woord (staan), te schande (maken), te beurt (vallen)

(3.11) go to water, go to work, pitch into, helping, to be on sale, speak to, to disgrace, to be given

The conclusion from these examples should therefore be that, apart from the examples in (3.8), 'te' does not seem interesting as preposition within a bare PP. On the other hand, the occurrences given with light verbs might be of interest to other researchers. For now, a further research into the aspects of 'te' would be beyond the scope of this thesis.

3.2.3 Activity Readings with *Op*

The preposition 'op' is not very high in the frequency table provided, but there is at least one reason why we should check this preposition: in English, it is reported by e.g. McIntyre that its direct English translation 'on' can be followed by any recording medium (e.g. 'on cd', 'on tape', etc.). For Dutch, we find even three categories following 'op': recording media (3.12), locations (3.13) and certain activities (3.14). In this section, I will go over these one by one.

(3.12) op + RECORDING MEDIUM: papier (paper), schrift (script), televisie (television), video (video), band (tape)

(3.13) op + LOCATION: tafel (table), straat (street), school (school), bed (bed), aarde (earth), kantoor (office), zee (see), stal (stable), zolder (attic)

(3.14) op + ACTIVITY: weg (way), bezoek (visit), reis (journey), stap (step), visite (visit), pad (path), komst (arrival), tournee (tour), zoek (search), vakantie (vacation), safari (safari), verlof (leave)

If we go over the nouns following 'op', we see that indeed, some recording media can be used bare after 'op'. However, the list in (3.12) above seems far less comprehensive than we would expect. But there is a reason for that: most of the texts of the Eindhoven Corpus are from the 1970's, a time in which some of the recording media (e.g. floppies, ZIP-drives, USB-disks, and the like) did not yet exist. The other nouns are readily translated with 'on' in English.

On closer scrutiny, however, 'television' seems to be the odd one out, as this is not a recording medium. It might be that 'television' should here be more like the activity of 'watching television'. In that case, 'television' actually is more like the nouns of the LOCATION-category in (3.12). Here we again find nouns that get an activity sense on their use without a determiner. In English, 'op' before these nouns is translated with 'at'. However, there are some special forms within this category: e.g. 'op straat' does not really denote an activity, and neither does 'op zolder'. And with some nouns (like 'stal'), the full form is ungrammatical.

In the last category, there also seems to be something strange at hand. We here find some timely interruptions ('vakantie', 'reis', 'verlof'), but not in any way like those that Stvan and this thesis identified as English N-based bare PPs (viz. 'breakfast', 'lunch', 'dinner'). English mostly uses 'on' for this kind of bare PPs, though some forms (like 'op stap', 'op komst' or 'op weg') seem to have no equivalent in English.² The category seems productive, however, as the corpus split test (as described in the introduction of this thesis) indeed yields an amount of types relative to the size of the corpus.

3.3 \exists -based bare PPs

In Dutch, 'met' ('with') and 'zonder' ('without') seem to appear quite freely with bare singular count nouns. Consequently, this forms an important difference between the Dutch and the

²Although it is questionable whether these are really 'count nouns' in the sense that 'on' gives them.

English data, as we already established in the previous chapter. There is, however, just as in English, an important caveat, which the reader might have already found in the frequency Table 3.1 above: while the negative 'zonder' appears far less frequent than the positive 'met', the former seems more likely to appear with bare singular count nouns. This section will go over some of the examples, and resolve whether 'met' and 'zonder' in Dutch indeed are to be compared to 'with' and 'without' in English.

To start off, with 'met', just as with 'with', we find that the preposition is used in a lot of distinct ways, certainly not in all cases of interest in our current research. However, contrary to English, with 'met', we are able to find at least some examples in which 'met' is followed by a bare singular count noun.

- (3.15) *Ikzelf stond gewoon voor de keus: zonder baard was ik precies
I myself stood just for the choice: without beard was I exactly
Pinocchio, met baard net Vincent van Gogh.
Pinocchio, with beard like Vincent van Gogh.*

"I just had to choose: without a beard I was just like Pinocchio, with a beard like Vincent van Gogh."

- (3.16) *Het wapen woog 4,4 kg (met bajonet 4,8 kg).
The weapon weighed 4.4 kg (with bajonet 4.8 kg).
"The weapon weighed 4.4 kg (4.8 kg including the bayonet)."*

- (3.17) *Je moet dan eerst laten zien dat er met overleg niets te bereiken
You must then first show that there with consultation nothing to arrive
valt.
is.*

"You will then first have to show that consultation will not get us anywhere."

However, one might object to some of the examples above. For starters, the bare appearance of 'met baard' ('with beard') in (3.15) might rely on the earlier bare use after 'zonder' ('without'), so that the utterer created a nice stylistic effect. Secondly, (3.16) seems to be an abbreviation. And concerning (3.17): is 'overleg' ('consultation') really a count noun here? For 'zonder', on the other hand, the examples provided by the corpus seems to be correct without doubt.

- (3.18) *Kan je hulp bieden aan een land dat tienduizenden gevangenen
Could you help offer to a country that tens of thousands prisoners
zonder proces gevangen houdt?
without trial improvised keep?*

"Could you offer help to a country that imprisons ten of thousands of prisoners without a trial?"

- (3.19) *Uiteindelijk is het niet herkenbaar tussen de andere huizen, zes stuks, alle
At long last is it not recognizable between the other houses, six pieces, all
onder één dak, zonder garage voor de Porsche en de kleine Cooper.
under one roof, without garage for the Porsche and the little Cooper.*

"At long last the house is unrecognizable between the other houses, six in total, all under one roof, without a garage for the Porsche and the little Cooper."

- (3.20) *Een dame op straat zonder hoed is geen dame.
A lady on street without hat is no lady.*

"A lady on the streets without a hat on can't be called a lady."

With these examples, we see that a proper English translation always need the indefinite determiner 'a' to complete the 'without'-part. Of course, it would be interesting to see whether

this always is the case, and hence whether English and Dutch are different with respect to \exists -based bare PPs. In the chapter on multilingual analysis, I will come back to this point. For now, it suffices to say that we established that indeed, Dutch allows for bare nouns following an \exists -based bare PPs, although the examples provided with 'met' are not beyond any doubt.

3.4 N-based bare PPs

As [Paenen, 2009] noticed, there are few N-based bare PPs to be found in Dutch. She started from the list of locations in [Stvan, 2007], translated these to Dutch, and found that there was ample evidence for a similar class of N-based PPs in Dutch. While I would certainly not argue otherwise, I do think we should have a full overview of the nouns appearing with more than one preposition in the Eindhoven Corpus, and not start from any pre-compiled list. I will therefore again use the method described for English in the previous chapter to look for all these nouns, categorize them, and report them below.

3.4.1 Locations

Concerning the locations, we indeed find a lot less nouns in Dutch which appear bare as the object of multiple prepositions. Only a quick look at Table 3.2 below reveals that the only clear candidates for appearing bare inside a PP form 'huis' and 'school' (respectively 'home' and 'school'). For the other nouns, there seem to be only a few prepositions with which the noun is able to acquire a grammatical status. For example: 'straat' ('street') only occurs frequently after 'op' (meaning 'at the street' or possibly 'on the streets', generic usage as in (3.20) above) and 'over' (meaning again 'at the street', but here only with introduced located objects). A similar line of reasoning is applicable to the other nouns from the table.

And still, when we go over the top two examples and compare these results to the English results in Table 2.2, is that for example 'school' is far less restricted in the amount of distinct prepositions it can form a grammatical bare PP with. An interesting difference, for example, is that 'in school' is grammatical in English, while in Dutch, this combination is out. Dutch here uses 'op' ('on') as the direct translation, which then again in English occurs far less frequent (though not ungrammatical).

Table 3.2: Nouns of the LOCATION category with (1) its English translation, (2) the number of occurrences of the noun in bare PPs, (3) the number of distinct prepositions the noun occurs with and (4) a specification of both columns: which preposition-noun combinations do we find and in what number? (Translations of the prepositions can be found in Table 3.1 above)

noun	tl	#	# d.p.	preposition (# of occurrences with noun)
huis	home	155	8	aan (3), bij (5), in (39), na (4), naar (79), tot (1), uit (2), van (22)
school	school	49	7	met (1), naar (10), op (25), over (1), per (2), tussen (2), van (8)
werk	work	38	7	aan (1), in (1), over (1), te (20), van (8), voor (4), zonder (3)
bed	bed	69	6	in (18), naar (34), op (11), per (2), uit (3), van (1)
zee	sea	30	6	aan (3), in (11), naar (2), op (4), over (1), uit (9)
arde	earth	16	6	aan (1), boven (1), naar (2), op (10), tussen (1), zonder (1)
tafel	table	52	4	aan (12), op (34), van (5), zonder (1)
straat	street	33	4	op (28), over (3), tot (1), van (1)
kantoor	office	13	4	naar (5), op (6), van (1), voor (1)
land	land	9	4	aan (2), te (2), tot (1), van (4)
wal	shore	4	3	aan (2), tussen (1), van (1)
boord	board	29	2	aan (24), van (5)

The interruptions that were identified by both [Stvan, 1998] and this thesis do not occur bare

with any preposition in Dutch. Other categories of N-based bare PPs were also not to be found, which became clear after categorizing all nouns appearing with more than five distinct prepositions in the Eindhoven corpus. For Dutch, we will have to conclude that this LOCATIONS-category is the only in its kind. And even in this category, the number of nouns that can appear with a multitude of prepositions seems limited, especially when we compare this number to English. Furthermore, the bare nouns can appear with only a limited number of prepositions, of which 'op' ('on'), 'naar' ('to') en 'aan' (again 'on', but then in the meaning of 'attached on') seem to be the most prominent.

3.5 Conclusion

In this chapter, we searched for bare PPs in Dutch. Again, we started off with P-based bare PPs, and again, we found that 'per' was on top of the frequency table. While 'per' in English was limited to the UNIT reading, in Dutch, we saw that 'per' can also take nouns of the MODE OF TRANSPORT and COMMUNICATION MEDIA class. In English, we saw that 'by' performed these functions. With 'op' then, we found that at least two types of nouns can follow this preposition without a determiner: RECORDING MEDIA and ACTIVITIES. It seems that this latter class has been largely neglected in the literature till now (although [van der Beek, 2005] does mention that nouns that denote a journey ('reis', 'vakantie') can appear after 'op').

Concerning \exists -based bare PPs, we found that Dutch allows bare singular count nouns to appear after both 'met' ('with') and 'zonder' ('without'). However, we found that for 'zonder', these examples were far more easy to find. Also, when we then translated these examples to English, we saw that for a natural translation, a determiner was needed. It might be worthwhile to research whether Dutch indeed is more like to have bare nouns following \exists -based bare PPs by using a multilingual corpus.

With regard to N-based bare PPs, we found that Dutch only seems to allow for nouns of the LOCATION-category to appear bare after several distinct prepositions. However, we see that the amount of prepositions that these nouns appear bare with is limited, especially in comparison with English. Furthermore, the amount of nouns in this class is much smaller compared to English.

Having analyzed all three forms of bare PPs in Dutch, we are now in a position to fill Table 3.3 with the Dutch data. Note that for some categories (e.g. 'onder' in Dutch or 'on' in English), I have not provided an in-detail analysis, but the reader is very welcome to check the statements on these specific parts.

The image that we acquire from this table is that English and Dutch are quite similar with respect to P-based bare PPs, apart from the BUSINESS AS MEDIA-category we found when analyzing the results for the English 'in'. For N-based bare PPs, English seems to be in the majority, while with \exists -based bare PPs, this picture is the other way around. Of course, it would be very interesting to further fill in the table for other languages. If we can indeed prove that, for example, categories of P-based bare PPs are to be found across every language, we might speak of bare PPs as a language universal. A similar argument can be made for N-based and \exists -based bare PPs.

To conclude the monolingual part of the thesis, let us go over three problems that we were not able to solve monolingual corpora. First of all, the monolingual analysis requires quite a lot of manual labor for each language, before we could finally come to a comparison. Of course, now that we have created a typology, in the future, this research will be far more guided. But in the next chapter, we will see whether statistical analysis of a multilingual corpus is suited to acquire fast and transparent results for this kind of contrastive analysis.

Another problem was that we have not yet been able to say anything about the semantics of certain P-based bare PPs. Especially 'per' seems interesting. In the first section of the fifth chapter, a section is devoted to finding a method to acquire a proper semantics for 'per'. A last disadvantage is that monolingual corpora do not seem to be able to enlighten us on the conundrum with 'with' and 'without': is it really so that English has a problem with the \exists -based

Table 3.3: A typology for bare PPs in English and Dutch. In the first column, we find the three types of bare PPs, in the second column, the categories of nouns (or preposition for \exists -based bare PPs), and the last two columns give the preposition English/Dutch uses for this form (for P-based bare PPs) or the number of occurrences (for N-based and \exists -based bare PPs). With regard to these last two columns: an asterisk (*) denotes that this form of bare PPs is far less used compared to the other language, a hyphen denotes the bare PP is absent in this language, and just as in 2.6, the 'numbers' (many, some and zero) are only indicative, and should not be interpreted with any real number.

P-based bare PPs	Category of nouns	English preposition	Dutch preposition
	N \Rightarrow UNIT	per	per
	MODES OF TRANSPORT	by/on	per/te
	COMMUNICATION MEDIA	by	per
	RECORDING MEDIA	on	op
	MEDIA AS BUSINESS	in	-
	CLOTHING	in	in
	ACTION	under	onder*
	ACTIVITY	on*	op
N-based bare PPs	Category of nouns	# in English	# in Dutch
	LOCATION	many	some
	INTERRUPTION	many	zero
\exists-based bare PPs	Preposition	# in English	# in Dutch
	<i>with</i>	zero	some
	<i>without</i>	some	many

bare PPs, or was it just the corpus that produced some faulty data? A multilingual analysis will certainly be of help here, and I will devote the second section of the fifth chapter to this question.

Part II

The Multilingual Perspective

Chapter 4

Statistical Analysis of the Multilingual Corpus

4.1 Introduction

While we have already taken a glance at the possibilities of a multilingual corpus in the first part of this thesis, it is now high time to discover and make use of the full potential of these data structures. Multilingual corpora, as for example [Johansson, 2007] notes, give us a bright new window on both the syntax and, for our purposes most importantly, the semantics of the linguistic phenomenon under study. In the now following, we will venture on a trip through multilingual corpora. On our first trip, we will see whether we can prove the hypothesis that English allows more bare nouns to appear as object of a preposition.

4.2 A Determined Hypothesis

In chapter 2, we saw that in English, there are quite a lot of bare PPs to be found. It is questionable whether this is also the case for Dutch and French, as we saw that, for example, the N-based bare PPs are far less common in Dutch. Our null hypothesis, nevertheless, would be that there is no difference between the three languages in their bare occurrences, and the typology given in Table 3.3 also points in that direction. However, using monolingual corpora to solve this matter has one important drawback: since the texts do not necessarily follow the same distribution in types and topics of texts (one features a section on fiction, the other does not, the one has a section on fashion, the other does not, and so on), it might be that we miss some important relations.

A solution would then be to use a **multilingual corpus** of translated texts. Here, in each language, the same meaning is expressed, and the only variable is the exact words that are used to infer the meaning. The multilingual corpus we employed is that of EuroParl ([Koehn, 2005]). It mainly consists of transcripts of meetings of the European Parliament. For every sentence, the speaker's native language is given.

However, this is the only information we get. The corpus is not tagged for part of speech, so, unfortunately, we can not start an analysis similar to the one in earlier chapters. Nonetheless, it might be that there is a way around. What we could do, is to follow these three simple steps:

- (4.1) Count the number of prepositions.
- (4.2) Count the number of prepositions followed by a determiner.
- (4.3) Divide the first by the second and see whether we find a (statistically significant) difference.

In counting the number of prepositions, I have chosen to only count those prepositions having at least fifty occurrences with bare nouns in the monolingual corpora. This requirement was posed due to some polysemy in the prepositions with a lower bare PP count. For example, 'te' in Dutch is only minimally used as a preposition ('te voet', 'te paard'), and far more frequently as particle of a verb.¹ I have added the regular expressions searched for in the three compared proceedings below.

(4.4) English: < (about | after | at | before | between | by | for | from | in | into | of | on | per | to | under | with | without) (a | an | the) >

(4.5) Dutch: < (van | in | op | met | tot | voor | naar | per | aan | bij | door | zonder | onder | uit | over | tussen | na) (een | de | het) >

(4.6) French: < (à | après | avant | avec | contre | de | d' | en | entre | par | pour | sans | sous | sur) (l' | le | la | les | un | une) >

The regular expressions have been searched for in the transcripts of the European Parliament meetings of the last quarter of 2000.² The number of hits of each of the individual steps are reported in Table 4.1 below.

Table 4.1: Frequency data of preposition directly followed by a determiner in English, Dutch and French using the 4th quarter of the 2000 proceedings in the Europarl Corpus, per daily proceeding. In the second column for each language, we find the total amount of prepositions. In the last column, these two counts are divided by each other, yielding the percentage of prepositions directly followed by a determiner.

Corpus	English			Dutch			French		
	# P+D	# P	%	# P+D	# P	%	# P+D	# P	%
ep-00-10-02	1044	3850	27,12%	1554	3177	48,91%	850	3414	24,90%
ep-00-10-03	735	2263	32,48%	1033	1910	54,08%	757	3114	24,31%
ep-00-10-04	3557	11899	29,89%	4935	10336	47,75%	2897	11320	25,59%
ep-00-10-05	1785	6324	28,23%	2625	5479	47,91%	1574	5828	27,01%
ep-00-10-06	276	812	33,99%	397	731	54,31%	255	816	31,25%
ep-00-10-23	1292	4515	28,62%	1871	3701	50,55%	1045	3893	26,84%
ep-00-10-24	4385	13807	31,76%	6125	11528	53,13%	2854	11385	25,07%
ep-00-10-25	4117	13687	30,08%	5892	11847	49,73%	3372	13212	25,52%
ep-00-10-26	1401	4905	28,56%	1937	4064	47,66%	1152	4678	24,63%
ep-00-10-27	390	1262	30,90%	483	1041	46,40%	401	1192	33,64%
ep-00-11-13	942	3113	30,26%	1369	2654	51,58%	708	2704	26,18%
ep-00-11-14	3887	12017	32,35%	5733	10836	52,91%	2890	11295	25,59%
ep-00-11-15	3653	12967	28,17%	5283	10985	48,09%	2955	11824	24,99%
ep-00-11-16	2148	7312	29,38%	3025	6316	47,89%	1633	6597	24,75%
ep-00-11-17	780	2383	32,73%	1006	2011	50,02%	590	2098	28,12%
ep-00-11-29	2983	9123	32,70%	4193	8039	52,16%	2611	9058	28,83%
ep-00-11-30	1747	5444	32,09%	2456	4828	50,87%	1501	5071	29,60%
ep-00-12-11	1284	4305	29,83%	1735	3568	48,63%	1012	3699	27,36%
ep-00-12-12	3615	11667	30,98%	5217	10250	50,90%	2645	10846	24,39%
ep-00-12-13	3648	11939	30,56%	4998	10192	49,04%	2690	10894	24,69%
ep-00-12-14	2200	6915	31,81%	2893	5976	48,41%	1622	5982	27,11%
ep-00-12-15	468	1773	26,40%	643	1444	44,53%	493	1750	28,17%
average	2106	6922	30,40%	2973	5951	49,79%	1659	6394	26,75%

¹For its logical counterpart, 'to', I expect that this phenomenon is less recurrent.

²This is an often used sample of the corpus. The sample consists of 1.079.512 words for English, 1.083.169 words for Dutch and 1.148.261 words for French.

From these results, it is clear to see that in Dutch, the percentage of prepositions followed by a determiner is much higher compared to both English and French. We could statistically verify this hypothesis using a paired test, as each individual corpus is conducting the same meaning. Since we can not assume normality, the non-parametric alternative has to be chosen. Using the Wilcoxon signed-rank test we find that $W(22) = 0, p < .01$, and so the null hypothesis of equal means has to be rejected. However, importantly, we also see that in French, there are even less determiners following a preposition than in English, $W(22) = 10, p < .01$.

While the former result is quite compatible with earlier observations (for example, the observation of [van der Beek, 2005] that only 'school' has similar behavior to the nouns defined by [Stvan, 1998]), the latter result seems to be off. There is no direct reason to presume that French seems to be less willing to have a determiner after the preposition, and by the reasoning proposed in the methodological section of this chapter would thus also have more bare PPs.

One should consequently ask himself whether the methodology used in this section was actually correct. I scrutinized the first 250 hits of the regular expression in (4.4) to try and answer this question. First of all, it is striking that we do not find a single 'real' bare PP as exemplified in the second and third chapter of this thesis. On the contrary, we find that slightly more than half of the prepositions are either part of a verb (exemplified in (4.7) and (4.8) below), idioms (4.9) or part of a larger construction (4.10).

(4.7) However, to my great surprise, I was **informed by a Belgian** (...)

(4.8) Please allow me **to propose**, on behalf of my group (...)

(4.9) Please allow me to propose, **on behalf of** my group (...)

(4.10) In words **with which** I identify completely (...)

So at least from this sample, we should conclude that there is a lot of pollution in our data. Such pollution of course devastates the truth value of the statistics given above. What is more, and might be even more of a strike to the methodology used above, is a point that we have yet not addressed: it might well be that English, Dutch and French have quite a different manner of dealing with several forms of determiners (e.g. articles, demonstratives and possessives). If that is the case, we introduce yet another unaccounted variable in our data, and of course, such an extra unaccounted variable only furhter questions the significance of our earlier result.

Consequently, it seems more fruitful to take another stance towards this multilingual corpus. In the next chapter, we will do just that, by performing statistical analysis not over the whole corpus, but an analysis per preposition.

Chapter 5

Multilingual Analysis per Preposition

5.1 Introduction

As we have seen in the previous chapter, a direct statistical analysis of the multilingual corpus now under study is hindered by the absence of part-of-speech-tags. Therefore, it seems sensible that we have to find other methods of analysis. In this chapter, I will consequently use more sophisticated methods than before to provide a more precise syntactic and semantic analysis per preposition with the help of the EuroParl corpus.

We will start off with the preposition 'per', which was on top of the frequency tables in both English and Dutch. Next is 'without', for which we failed to give a clear analysis in English only.

5.2 Peculiarities of *per*

In earlier sections on 'per', we found that its main semantics is to transform the noun into a UNIT measure: 'per day' becomes 'each day', 'per region' becomes 'each region'. This UNIT usage of 'per' seems to be available in most languages. In German, for example, one uses 'pro + N' to acquire similar readings. In French, the UNITer is 'par + N'. However, in the chapter on Dutch, we found that the Dutch language employs 'per' for also performing some other functions. In this section we will therefore see whether we are able to give a proper semantics for this preposition, which seems to have its origin in Latin.¹

5.2.1 Method

In this section, I again use the EuroParl multilingual corpus of proceedings of the European Parliament as the primary dataset. Again, only the last quarter of the 2000 proceedings is used, so again, the corpus is about 1 million of words in size. In both English and Dutch, the occurrences of 'per' are then searched for.² Each of the results and its accompanying translation is stored in a separate file. For the original language's sentences containing 'per', 'per' and the noun after it are kept, the rest of the sentence is deleted. In the translated sentences, the translation of 'per + N' are then manually looked for and extracted.

¹Also, one might be interested why 'per' survived as a preposition, while other Latin prepositions have become archaic and are only used within idioms ('ex aequo', 'in spe' to name a few). However, such an analysis is way out of the scope of this thesis, and will therefore be left as a topic for further research. A good starting point here might be a diachronical corpus, which seems yet to be under development for Dutch.

²This time the regular expression for searching the corpus is not that hard: simply search for '<per>' in both languages.

To ascertain the specific semantics of ‘per’ in both English and Dutch, we are of course most interested in those cases in which ‘per’ is not readily translated, but another construction is used to convey the same meaning. These examples will consequently be of special interest to us to acquire an idea of a semantics for this preposition.

5.2.2 Results

The (regular expression) search for ‘per’ in the corpus yields 103 matches in the English translation of the proceedings and 196 results for the Dutch translation. Of these matches, we find that 54 occurrences of ‘per’ are readily translated into the other language. As has been stated in the methodology section above, we are most interested in those examples which are *not* readily translated by ‘per’. In this way, we are able to find interesting aspects of this preposition, which might eventually help to set up a proper semantics.

The Dutch Data

As Dutch yields the most matches, we will first look at these constructions and their translations. In Table 5.1 we find the data categorized by (kind of) translation into English. In the next paragraphs, I will refer to each class by its Roman numeral found in the first column of this table.

Table 5.1: Frequency data for translations of ‘per’ in Dutch into English. In the last column, a typical example of the defined category is given. The Roman numeral in the first column will be used in the text to further work out the specific content of each class. The horizontal line in the middle of the table divides the more productive categories from the idioms, exceptions and non-translated forms of ‘per + N’.

RN	Dutch	English	#	Typical example
I	per	per	54	per persoon ⇒ per person
II	per	a	24	per week ⇒ a week
III	per	ADV/ADJ	18	per jaar ⇒ annually
IV	per	by	17	per spoor ⇒ by rail
V	per	each/every	14	per lidstaat ⇒ each Member State
VI	per	N-by-N	10	per thema ⇒ subject by subject
VII	per	from N to N	6	per land ⇒ from country to country
VIII	per	N basis/based	3	per activiteit ⇒ activity-based
IX	per	idioms	21	per slot van rekening ⇒ after all
X	per	exceptions	16	per november 2000 ⇒ as of november 2000
XI	per	not translated	12	

What might be the most striking in this table is the occurrence of ‘a week’ as a translation of ‘per week’. What might a determiner be doing there? A few examples (below, in (5.1), (5.2) and (5.3)) however suffice to see that ‘a’ here might actually *not* function as a determiner.

(5.1) *In een wereld waarin de bevolking met 80 of 90 miljoen per jaar toeneemt, moet er een markt zijn voor Europees voedsel.*
 In a world where the population with 80 and 90 million per year accumulates, must there a market be for European food.

“In a world where the population is growing by 80 and 90 million a year, there must be a market for European food.”³

³In all the examples in this section, both the original and the translated version of the sentence are copied exactly from the corpus. The glosses in the second line are mine.

- (5.2) *Naar mijn weten bedraagt de totale handel tussen de EU en de VS*
 To my knowledge amounts the total trade between the EU and the US
per dag ongeveer 1 miljard euro.
 per day about 1 billion euro.

"I think I am right in saying that EU/US trade amounts to about EUR 1 billion **a day**."

- (5.3) *Ik wijs u erop dat dit centrum per dag gemiddeld 800 tot 1.000*
 I point you to that this centre per day averagely 800 and 1.000
vluchtelingen opvangt.
 refugees receives.

"I should like to inform you that this centre processes an average of between 800 and 1.000 refugees a day."

What we do see, on the other hand, is that there is a common construction in all these examples. When looking at the data more closely, we see that 'a' as a translation of 'per' most of the time occurs with periods of times (days, weeks, months and years), but even there seems to be limited, as (5.4) should show. Only in a minority of the cases (3 of the 23 examples attested), other nouns can be plugged into this construction. Two of these three examples can be found in (5.5) and (5.6).

Furthermore, the construction 'a + N' only seems grammatical when preceded by both a number and a mass noun (as in 5.1 and 5.2 above) or a plural noun (see 5.3). This heavily limits the usage of 'a' as a sort of substitution for 'per'.

- (5.4) *X minutes an hour, X hours a day, X days a week, X days a month, X weeks a year, #X years a decennium, # X years a millennium.

- (5.5) *Wij hebben in de afgelopen 18 maanden de prijzen voor ruwe olie zien stijgen*
 We have in the last 18 months the prices for crude oil seen rise
van USD 9,75 tot USD 33 per barrel.
 from USD 9.75 to USD 33 per barrel.

"Over the last 18 months we have seen the price of crude oil rise from USD 9.75 to USD 33 a barrel."

- (5.6) *Afvalvernietiging kost 75 pond per ton.*
 Garbage disposal costs 75 pounds per ton.

"They now cost €75 a ton for disposal."

A hypothesis here might be that the construction 'a + PERIOD OF TIME' is likely to be a remnant of the French preposition 'à', which lost its accent-grave along the years. A Google search⁴ revealed that French indeed sometimes uses this construction ('sept jours à semaine', a common children's song about the weekdays, is a primary example, 'seven days a week'). However, a closer look with the help of the multilingual corpus is necessary to establish this relationship, and is, because of my limited knowledge of French, outside the scope of this thesis.

In category III we find 'per + N' not translated by a construction with a preposition, but with an adverb or an adjective. In these cases, we see that the noun N is suffixed by '-al' or '-ally', yielding a meaning similar to 'per N'. Some examples can be found in (5.7), (5.8) and (5.9)). While Dutch also has this possibility of suffixing a noun to get an adverb, it is interesting too see that this behavior seems to be far less common there. For example: we do not find a single example in the now attested corpus of an appearance of 'per N' in English translated by an adverb in Dutch, as is clear from Table 5.3 below.

⁴With the kind help of Marten Postma.

- (5.7) *Tot slot sta ik positief tegenover het idee om één keer per jaar bijeen te komen om de voortgang van het programma te bekijken.*
 To close stand I positive against the idea of one time per year together to come to the implementation of the programme to review.
 "Finally, I welcome the idea that we should review the implementation of the programme **annually.**"
- (5.8) *Ten derde en ten slotte moet het actieplan en de uitvoering ervan per regio worden bekeken.*
 To third and to close must the action plan and the way of implementation of it per region be observed.
 "The action plan and its way of implementation must be observed **regionally.**"
- (5.9) *Bovendien kan de sector van de digitale media de komende 10 jaar met 20% per jaar groeien.*
 Furthermore could the sector of the digital media the coming 10 year with 20% per year grow.
 "Furthermore, the digital media sector could achieve 20% **annual** growth over the next decade."

As for category IV, the 'per + N' in Dutch translated by 'by + N' in English, we see that we find here the class of 'per + MEANS OF TRANSPORTATION'. An example can be found in (5.10) below). Strangely enough, we do not find the class of 'per + MEDIUM', mentioned in the introduction of this section. It is not exactly clear why this is the case; it might be that these cases are not particularly common in political texts like the proceedings of the European Parliament. In more than half of these examples found (9 of the 17) 'by' is used to translate the idiomatic 'per definitie', two examples of which are shown in (5.11) and (5.12).

- (5.10) (...) *het vervoer van gevaarlijke stoffen per spoor* (...)
 (...) the transport of dangerous goods by rail (...)
 "(...) the transport of dangerous goods **by rail** (...)"
- (5.11) *We zijn hier allemaal per definitie democraten.*
 We are here all per definition democrats.
 "We are all democrats here **by definition.**"
- (5.12) (...) *misdadige praktijken, die per definitie supranationaal zijn.*
 (...) criminal acts that per definition supranational are.
 "(...) criminal acts that are supranational **by nature.**"

In category V we find examples of 'per' which are translated using quantifiers like 'each' (in (5.13)) and 'every' (5.14), but also other individuating expressions, as (5.15) shows.

That 'per' is translated with the help of these quantifiers might, in return, help us with finding a semantics for 'per'. Importantly, while 'each' and 'every' are close relatives, there is a subtle difference in meaning between the two.

In her PhD thesis, [Tunstall, 1998] states that "a sentence containing 'each' can only be true of an event which has a totally distributive event structure, where each individual object in the restrictor set of the quantified phrase is associated with its own subevent, and all the subevents are differentiated on some relevant dimension." In other words, 'each' is strongly individuating. Every is then "subject to the weaker requirement that there be at least two different subevents."

In the current sample, 'each' as a translation of 'per' outnumbers 'every' by 12 to 2. While this might not yet say anything due to the small size of the sample, this indicates that the meaning of 'per' is closer to that of 'each' than to that of 'every'. Scaling up the corpus to not only the last

quarter, but the whole 2000 proceedings, yields again an advantage for 'each': 58 to 30.⁵ This is solid evidence for our claim that 'per' should be more like 'each' in its semantics.

(5.13) *Momenteel vallen er in Europa ongeveer 50.000 verkeersdoden per jaar.*
At the moment fall there in Europe approximately 50.000 traffic deaths per year.

"At present in Europe, we have approximately 50.000 deaths on the road **each year**."

(5.14) *Mijn indruk was dat er bijna per dag vooruitgang was.*
My impression was that there almost per day progress had been.

"I had the impression that there had been progress almost **every day**."

(5.15) *Het is zeker gemakkelijker om per geval pragmatische besluiten te nemen (...)*
It is certainly easier to per case pragmatic decisions to take (...)

"It is certainly easier to make pragmatic decisions **in individual cases** (...)"

In category VI, we meet an old friend in the N-by-N cases, which are exemplified below. We quickly went over these cases in the section on 'by' in the second chapter. Here, we noted that while these forms might be related to bare PPs in syntax, they do not seem to have any semantic relationship with them. While this conclusion seemed justified for 'by' itself, we see that 'per + N' actually does have a semantic relative in the 'N-by-N' construction.

As for the conclusions we can draw from these translations, [Jackendoff, 2008] notes that "the main sense of 'N by N' is of some sort of succession". 'N by N' denotes something like 'one N at a time'. Again, this translation states that 'per' should have a strongly individuating sense.

(5.16) *Daarom moeten wij per land over een exacte inventarisatie beschikken.*
So must we per country about an exact overview have.

"So we must have an exact overview of the situation, **country by country**."

(5.17) (...) *dat het de voorkeur verdient om de opleidingen elk jaar of per thema op een andere politieacademie in een andere regio van de Unie te organiseren.*
(...) it would be better if the courses each year or per subject at another police academy in an other region of the Union to organize.

"It would be better to take everybody to one place or to have courses which move **year-by-year or subject-by-subject** to different police colleges and academies in different parts of the Union."

(5.18) *Ik ben vast van mening dat per geval bekeken dient te worden of de overstap van eenparigheid naar stemmingen met gekwalificeerde meerderheid gemaakt dient te worden.*
I firmly believe that per case considered has to be whether the change from unanimity to QMV made has to be.

"I firmly believe that any change from unanimity to QMV should be considered **on a case-by-case basis**."

Category VII is home to yet another special construction: here we find forms of 'per N' translated as 'from N to N'. We only find this happening with 'per land', as (5.19) shows.

⁵Here, I used a parallel search for 'per' in Dutch and 'each' or 'every' in the English translation. However, I did not check each example individually, so it might be that 'each' or 'every' in the hits might be used for translating another word in the sentence. However, given that 'every' is slightly more frequent than 'each', we are on the safe side when claiming that 'each' is more common as an alternative for 'per'.

(5.19) *Deze cijfers zijn echter misleidend omdat zowel het concept als de dekking van het stelsel per land sterk verschilt.*
 The figures are however misleading since both the concept and the coverage of the system per country significantly differs.

“However, these figures are somewhat misleading since both the concept and the coverage of the system differ significantly **from country to country.**”

In category VIII we see the cases in which ‘per’ is translated with the help of ‘base’ or ‘basis’. We already saw such an example in (5.18) above.

The other three categories form respectively idioms (21 tokens, not really of interest here), some exceptions in which ‘per’ is translated by yet another form (16 tokens) and some untranslated forms (12 in total). For completeness, I have the idioms and exceptions and their respective translations in Table 5.2 below.

Table 5.2: Idioms and exceptions in translating to English from Dutch.

Dutch original	English translation
per 1 juli 2001	from 1 July 2001
per 1 april aanstaande	next April
per abuis	the wrong way
per begrotingsterrein	at sector level
per brief	in a letter
per bus	buses
per categorie	of various content categories
per categorie	of components
per definitie	deep down
per definitie	must be
per definitie	be aware that you are also
per diersoort	species-specific
per direct of op korte termijn	the immediate or soon to take place
per kassa	for both cash and off-balance-sheet items.
per la rinfondazione comunista	the Communist Refoundation Party
per november 2000	as of November 2000
per ongeluk	unfortunate enough
per ongeluk	an oversight
per saldo	on balance
per saldo	at the final count
per schip	for the ship
per se	have to
per se	the need
per se	in essence
per se	at all costs
per sector	on all the individual sectors
per slot van rekening	after all (5x)
per slot van rekening	no more
per spoor	for the railway.
per spoor	as railways
per type	according to the type
per vrachtwagen	for the lorries
per weg	road transport

The English Data

Table 5.3: Frequency data for translations of 'per' in English into Dutch. In the last column, a typical example of the defined category is given. The Roman numeral in the first column will be used in the text to further work out the specific content of each class. The horizontal line in the middle of the table divides the more productive categories from the idioms, exceptions and non-translated forms of 'per + N'.

RN	English	Dutch	#	Typical example
I	per	per	54	per person ⇒ per persoon
II	per NUMBER	procent/promille	34	75 per cent ⇒ 75 procent
III	per	ieder/elk	2	per Member State ⇒ iedere lidstaat
IV	per	exceptions/idioms	9	per diem ⇒ op dagbasis
V	per	not translated	4	

With English, we see that more than half (54 of 109 yields 52%) of the examples with 'per' can be readily translated into Dutch. The almost only non-'per'-translation category is the one consisting of 'per' followed by a 'number'⁶: 'per cent', 'per thousand' and 'per million' are here the typical examples. In Dutch, these expressions are translated by 'procent', 'promille' and 'op één miljoen'.

We can thus conclude that Dutch uses 'per' far more frequent than English for the sense that denotes a UNIT reading. For the sake of completeness, I have added the exceptions and idiomatic forms in the table below, just as we did for Dutch. We again find that 'per se' is translated differently. Also, the use of 'as per' is unexpected. This construction can appear also appear with a determiner, as is shown in Table 5.4 below.

Table 5.4: Idioms and exceptions in translating to Dutch from English.

English	Dutch
as per 1 January 2001	op 1 januari 2001
per diem	op dagbasis
per Member State	'één commissaris/ één lidstaat'
per se	weliswaar
per se	als dusdanig
as per the agreement	Zoals het Parlement en de Raad waren overeengekomen
as per the schedule	volgens het toen overeengekomen tijdschema
as per the common position	conform het gemeenschappelijk standpunt
per view	'pay-per-view'

5.2.3 A short note on 'per + MEANS OF TRANSPORTATION'

A special category of bare singular count nouns that can appear after 'per' in Dutch are the modes of transportation, an example of which we found in (5.10) above. In the previous section, we argued that all of these examples are translated with 'by' in English. However, further research into this category of nouns by Postma (p.c.) in French yielded an interesting result, which might have some impact on our analysis so far. In this paragraph, I will thus quickly go over Postma's examples and conclusion, and will then see whether his suggestions for French might also be compatible with English.

Postma starts off from the observation that 'per + MEANS OF TRANSPORTATION' has an alternative in 'met + DET + MEANS OF TRANSPORTATION' in Dutch. In French, we find a similar

⁶Number between parentheses: 'cent' is, of course, not actually a number.

distribution: there is 'en', which always has a bare object, and there is 'par', which needs a determiner to be grammatical. Such a division of labor without clear reason is weird from a point of view in which language is optimal, i.e., a language is optimally suited for both speaker and hearer to convey different meanings with different forms.

Postma therefore set out to analyze all Dutch-French examples in the EuroParl corpus in which either 'per' or 'met' was followed by a mode of transport.⁷ The list of these was compiled using the Dutch Wikipedia article on means of transportation.⁸ He found that indeed, French uses 'en' and 'par' as translations, but when scrutinizing the French examples more closely, he was able to make a far stronger conclusion: French employs 'en' when persons are to be transported, while 'par' is the preferred form when goods are transported. This interesting conclusion should of course be backed up some evidence. Some examples are provided below:

(5.20) *Monsieur le Président, j'espère que les deux orateurs qui devaient me précéder ne venaient pas en train et n'ont pas été retardés.*
 Mr the President, I-hope that the two members that will me precede not came not by train and not-have not been delayed.
 Mr President, I hope that the two members due to speak before me were not travelling **by train** and have been delayed.

(5.21) *Monsieur le Président, tous ceux qui sont arrivés ici en voiture, en train ou en avion ont pu constater l'ampleur des dégâts causés.*
 Mr the President, all those who have arrived here by car, by train or by aeroplane have can see the-extent of-the damage caused.
 Mr President, everyone who has travelled here **by car, train or aeroplane**, has been able to see the extent of the damage.

(5.22) *Il est très difficile de transporter un vélo par le train.*
 It is very difficult to transport a bike by the train.
 It is very difficult to transport a bike **by train**.

Here, we see in both (5.20) and (5.21) that when people are transported, 'en' is used. But in (5.22), goods are transported, and then French uses 'par' followed by a definite article. In Dutch, we do not see such a distinction: in both (3.4) and (3.5), we see that 'per' is chosen, even when the subject of the transport are humans. In French, however, the evidence is overwhelming: in 90% of the cases, French follows the given pattern.

What are then the deviations to this rule? Postma shows that the most important one is the following: when an expression is modified, French again resorts to 'par'. This becomes clear from the following example:

(5.23) (...) *parce que le personnel, tout le monde le sait, doit rentrer par le train de 15 heures.*
 (...) because the personnel, all the world that knows, must return by the **train** of 15 hours.
 "(...) because the personnel, as the whole world knows, must return on the 15 o'clock train."

Back to English. While it is clear from the above examples that the 'persons vs. goods' distinction does not hold in English (as 'by' is used as a translation in every single example), for modification, 'by' cannot be used. From the example given, we see that English often resorts to 'on' as the alternative. In the EuroParl corpus, modification of modes of transport is scarce, but a second example, next to (5.23), is provided in (5.24) below.

⁷For 'met', of course, the mode of transport needed to be accompanied by a determiner.

⁸Found on <http://nl.wikipedia.org/wiki/Vervoermiddel>

- (5.24) *Hij is per trein uit Parijs onderweg.*
 He is by train from Paris under way.
 "He is on his way here **on the train** from Paris."

An additional Google search however confirms our hypothesis that 'on' indeed is used when the mode of transport is modified, no matter what the mode of transport is. It thus seems that 'on' is the alternative for 'by' when a nominal is modified. This might be an interesting point to start off for further research: do we find such relations also with other kinds of bare PPs? A clear starting point would be 'per + MEDIUM' in Dutch, which has an alternative form in 'via + MEDIUM'.

5.2.4 Conclusion

In this section we have seen that 'per' in Dutch is used and preferred for a multitude of functions for which English prefers other forms. Working out the details of these other forms have helped finding a more concrete semantics for 'per'. Especially the use of the quantifier 'every' gives good evidence for a strongly individuating semantics. Also, the use of expressions like 'USD 33 a barrel' were not at all expected, but do give us insight in the peculiarities of 'per'.

During the research on 'per', I stumbled upon some further questions. In the following paragraph, I will give these as suggestions for further research.

5.2.5 Suggestions for further research

A road which we have not taken in this section, is that of diachronic research. As we have seen, 'per' is in many ways the odd one out under the prepositions. The preposition is, first of all, never followed by a determiner. Secondly, it opens multiple slots. We have seen so in Dutch: 'per' opens not only a slot to turn the whole construction in a UNIT, but also can license the bare occurrence of nouns in the MEANS OF TRANSPORTATION or MEDIUM class. In English, we saw that 'by' is reserved for the latter purposes. One could question two things: when and why did 'per' appear in the Dutch language, and when and why did 'per' get the extra functionality, that English reserves for 'by'? A diachronic research might be of help here.

In the previous, I have also left some issues of ambiguity aside. One could ask oneself whether the two uses of 'per', as a so-called UNITer or the opener of a slot for nouns of the MEANS OF TRANSPORTATION-class could, in one sentence, be intertwined. Here, I would argue, that this could indeed be the case. In (5.25) we indeed find such an example.

- (5.25) Er komen vier mensen per auto.
- a. Four people will come by car ($\rightarrow |cars| = \{1, 2, 3, 4\}$)
 - b. In each car, there will be four people ($\rightarrow |cars| \leq 2$)

- (5.26) Er komen vier mensen per fiets.
- a. Four people will come by bicycle.
 - b. # On each bicycle, there will be four people.

Is one reading preferred, or does the context decide which reading is the right one? Whatever might be the case, it is clear that (5.26b.) is out. However, this puzzle might be more of a pragmatic one than a semantic one, as it seems to have a relation our world knowledge.⁹ Because of this, we will leave this question aside in this thesis. In the next section, we will once again return to the main subject of this thesis: syntactic and semantic problems with bare PPs. This time we will scrutinize the preposition 'without'.

⁹Another puzzle comes with "U moet per SMS betalen" (typically translated as "You have to pay every/by SMS"), which could become ambiguous when paying by SMS will become more familiar over time.

5.3 Without Revisited

In this section, we will shift our focus to \exists -based bare PPs. In most languages, the prepositions denoting a meanings similar to those of 'with' and 'without' license the occurrence bare nouns. For example, [Le Bruyn et al., 2009] state that in both French and Dutch, a noun can appear bare quite freely after both these prepositions. However, in the monolingual analysis of chapter 2, we found that in English, it seems to be far less acceptable to leave out the determiner with both 'with' and 'without'. In the subsequent chapter, we saw that in Dutch, indeed, 'zonder' ('without') can be followed by bare nouns. For 'met' ('with'), however, these occurrences were found to be far less frequent.

There are at least two reasons why we would not expect this behavior. First of all, English and Dutch are similar in article usage, as for example [de Swart and Zwarts, 2009b] show. If this is the case, there is no reason why neither 'with' nor 'without' are rarely found with bare singular count nouns. Secondly, there is no obvious difference in meaning between the bare, the definite and the indefinite form with these PPs: at least in Dutch, the bare form 'zonder hoed' ('without hat') is not in any way semantically different from the indefinite 'zonder een hoed' or the definite 'zonder de hoed' ('without a/the hat'). We would thus expect to find no clear differences in the number of appearances of bare and full nominals. If there are differences, however, we might question the established fact (see [de Swart and Zwarts, 2009b]) that English and Dutch are quite similar in their usage of articles.

To find a way out of this conundrum, the multilingual corpus employed throughout this thesis might again be helpful. In this section, I will try to prove that indeed, English is different from Dutch with respect to the \exists -based bare PPs. We will do so by searching the corpus for every occurrence of 'without', and checking how the definite, indefinite and bare NPs following this preposition are translated into Dutch. In the methodology section below, I will first of all work out the methodology of this specific corpus study. In the section thereafter, I will show the results of the study and provide some analyses. The outcomes of the study are then summarized in the conclusion of this section, and then form the start of a new corpus experiment.

5.3.1 Method

The method employed in this section is quite similar to that used with 'per'. Again, the same subsection of the multilingual EuroParl corpus is employed, existing of the proceedings of the last quarter of the year 2000, consisting of just over a million words in both English and Dutch. However, with 'with' and 'without', there are some important caveats. With 'per', we were certain that every data point was of interest: 'per' appears almost exclusively followed by a bare singular count noun. For 'with' and 'without', this surely is not the case. First of all, both prepositions can appear with both singular and plural noun phrases, either with or without a determiner. Secondly, 'with' and 'without' have some other senses, which are far less relevant under these circumstances. For example, 'with' is often the particle of a verb ('compare with', 'in line with' etc.) and its counterpart 'without' often introduces a relative clause. Some examples of this latter behavior are provided below.

(5.27) Treaties are not rewritten **without affecting existing legislation**.

(5.28) We know that the Commission discussed this question at its last meeting, **without coming to a decision**.

(5.29) **Without action being taken from the Greek side** we will pursue the procedure already underway.

So, when searching the corpus, we have to expect quite some pollution of the statistics. Since the multilingual corpus is not tagged for part of speech, it is not easy to automatically (i.e. with a regular expression) neglect cases as in (5.27) above. What we could do, on the other hand, is to manually categorize all the hits. However, for 'with', this is bound to take a lot of time: in

the monolingual Brown Corpus, we had over 7000 hits. Luckily for us, ‘without’ is a far less common preposition. We should thus focus here on ‘without’, and presume that our results are to be extrapolated to ‘with’.¹⁰

In the multilingual EuroParl corpus we will thus search for all of the occurrences of ‘without’. We will manually filter out the uses in which ‘without’ is the head of a relative clause or is otherwise not followed by a noun phrase. Of these occurrences of ‘without’ followed by a noun phrase, we then select the singular count nouns, either bare or preceded by a (in)definite determiner. For each of these hits, we acquire the Dutch translation from the corpus, and then categorize each hit of ‘zonder + NP’¹¹ into one of three ‘article’-categories (definite, indefinite or bare) and a ‘residue’-category into which we add the translations which do not consist of ‘zonder + NP’.

The counts of these categories will give us a frequency matrix which we can use for drawing a much longed-for conclusion. The null hypothesis here, of course, should be that ‘without’ and its Dutch equivalent ‘zonder’ are equal in their use of accompanying determiners. In the next section, we will see whether this null hypothesis holds, or whether we are able to falsify it.

5.3.2 Results

Searching the corpus with the regular expression for ‘without’ yields a total of 639 hits. Of these 639, we find that 258 (40,4%) of these are not followed by a noun phrase, but rather form the introduction of a relative clause. For the remaining 381 hits, I assessed whether the noun in the noun phrase following ‘without’ was a singular count noun, using the method of Kiss, explained in the introduction of the second chapter of this thesis. Here, I did not delete modified noun phrases, since I find that these are not syntactically different from the unmodified forms. This procedure yielded 129 hits, 20,2% of the total occurrences.¹²

These remaining 129 hits can then be categorized in three categories, according to their (lack-
ing of a) determiner: definites, indefinites and bare nouns. Definites here are defined as (possibly modified) nouns preceded by the determiner ‘the’ or ‘this’. Indefinites are either preceded by ‘a(n)’ or ‘any’, and bare nouns, of course, do not have a determiner. For each hit, the Dutch translations were also categorized according to these three categories. For the cases in which ‘without + NP’ was translated by an adverb (e.g. ‘without a break’ ⇒ ‘ononderbroken’) or other constructions (e.g. ‘without exception’ ⇒ ‘bijna alle’ (‘almost all’)), the hit was placed in an ‘residue’-category.

This procedure then yields Table 5.5 below.

Table 5.5: Frequency table for translations of ‘without’. On the vertical axis, the English category is displayed, on the horizontal axis, we find the Dutch categories.

ENGLISH / DUTCH	# bare	# indef	# def	# residue	# total
# bare forms	16	3	0	46	65
# indefinites	10	17	1	24	52
# definites	3	1	6	2	12
# residue	32	14	6	-	52
# total	61	35	13	72	181

¹⁰However, monolingual research in the Brown Corpus did give us some indication that we cannot totally equalize the two prepositions. But since ‘without’ was far more likely to have bare singular count nouns as its complement, I would argue that research into ‘with’ will not yield to different results compared to those reported in the following sections.

¹¹For my non-Dutch readers: ‘zonder’ is the direct Dutch translation of the English ‘without’.

¹²When one would compare this number to the frequency reported in Table 2.1 of chapter 2, one could be surprised. However, it is important to once again stress that the frequencies reported in the table there are only indicative. One notices this when one looks at the data more closely: one then sees that most of the bare nouns following ‘without’ are actually part of a plural compound. In this corpus, this is also the case, and these hits are then excluded from the data.

What conclusions can we draw from this simple frequency table? Well, first of all, it is clear that under normal circumstances, bare forms are translated with bare forms, indefinites with indefinites, and definites with definites. But there are some interesting deviations to this rule. When translating English 'without + indefinite NP' to Dutch, we see that 10 out of 52 hits (19,2%) are translated with a bare noun. The other way around, an English bare form to a Dutch indefinite, we see that this is only 3 out of 65 (4,6%).

We could thus hypothesize that Dutch is allowed to use 'weaker' (in the sense of introducing discourse referents: bare nouns do *not* introduce discourse referents, indefinites *introduce* discourse referents and definites *refer* to discourse referents) forms than English in the company of \exists -based bare PPs. To prove this hypothesis, we should first of all categorize our results. First of all, the cases in which the Dutch uses either a bare form or an indefinite to represent a definite in English, and the case in which Dutch uses a bare form to represent an indefinite, can be grouped together as 'weaker forms'. Secondly, the 'stronger' forms are those in which Dutch uses a definite as a translation of an English indefinite, and an indefinite or definite to translate a bare form. And lastly, the cases in which both the original and the translated text use the same form can be grouped together. I believe that a graphical representation will surely make the last paragraph more clear, so I have added a graphical overview of the last paragraph in Figure 5.1 below.

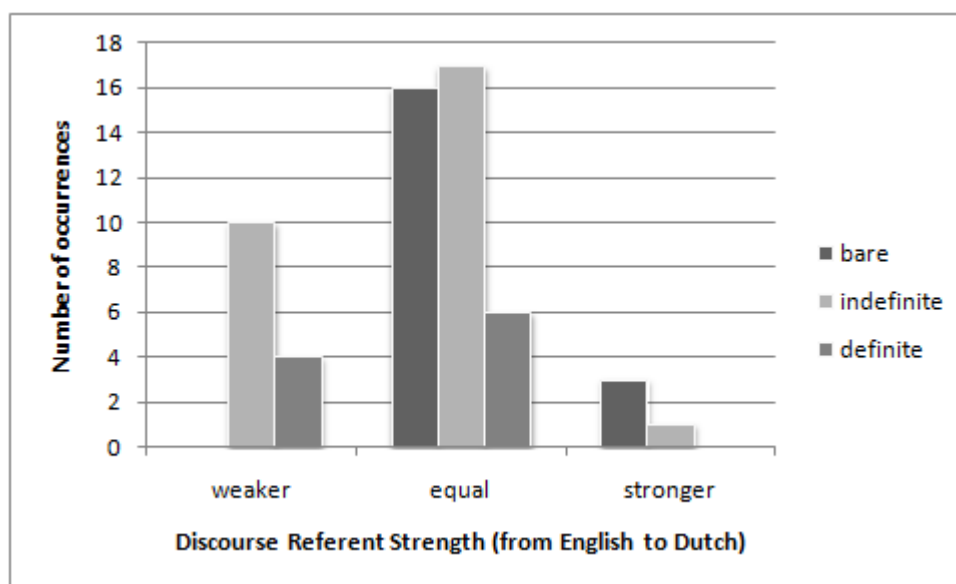


Figure 5.1: Number of occurrences per change in discourse referent strength, divided into the three categories of articles. Note that for bare forms the column 'weaker discourse referent strength' is necessarily empty, as bare nouns are essentially the weakest forms. A similar remark (but the other way around) must be made for definites and the category 'stronger discourse referent strength'.

We are now able to do an analysis of variances (ANOVA), with the number of occurrences as the dependent variable, and the discourse reference strength of the translation (either weaker, equal or stronger) as the independent variable. But before performing this analysis, we have to test for the homogeneity of the variances, since otherwise a parametric ANOVA should not be used. Since Levene's test yields an insignificant result ($L(2,6) = 3,64, p > .05$), we are entitled to use an ANOVA.

Doing the ANOVA-test yields $F(2,6) = 5,27, p = 0.048$, and the statistical test is thus significant on an α of .05. We have thus falsified our null hypothesis that English and Dutch are similar in their use of articles after 'without'/'zonder'. When we then apply post-hoc testing (Tukey Test) to find out which of the differences between the groups was found to be significant, we find that there is only a significant difference between the 'equal discourse reference strength' and the 'stronger discourse reference strength'. This would imply that English is more likely to use a

stronger discourse referent ('a' and 'the' instead of a bare noun, 'the' instead of 'a'), but *not* the other way around: Dutch is not more likely to use a weaker discourse referent (Dutch does not significantly use a bare noun where English uses a determiner).

As previous chapters have shown that it is actually worthwhile to go over various separate examples to verify the statistical results (as we saw in chapter 4), in the next subsection, I will present some examples in each category, to get a clearer look on what comprises Table 5.5 above.

Bare Forms

With the English bare forms, most translations in Dutch end up in the 'residue'-category. A large sample of these is taken up by adverbs for 'without delay' and 'without doubt', as examples (5.30) and (5.31) show. 16 of the remaining bare nouns nicely translate into a bare noun, of which (5.32) is an example. With only three cases, we see that Dutch prefers a indefinite noun as a translation. These three are copied into (5.33) and (5.34) below.

(5.30) without delay ⇒ onmiddelijk, onverwijld, (pijl)snel

(5.31) without doubt ⇒ ongetwijfeld, zeer zeker, spreekt vanzelf

(5.32) *Het gaat om het recht om zonder beperking toegang te krijgen tot documenten, (...)*
 It concerns the right of without restriction access to get to documents, (...)
 documents, (...)

"It concerns the right of access to documents **without restriction**, (...)"

(5.33) *Dit is wat wij zonder enige aarzeling en zonder enig uitstel moeten doen, en wel op massale schaal.*
 This is what we without any hesitation and without any delay should do, and yet on massive scale.

"It has to be done, **without hesitation** and **without delay**, and on a massive scale."

(5.34) (...) *dit eerste pakket richtlijnen over de veiligheid op zee vormt zonder enige twijfel een van de belangrijkste daden die ons Parlement in het lopende jaar 2000 heeft gesteld.*
 (...) this first package directives on the safety at sea forms without any doubt one of the most important actions that our House in the current year 2000 has posed.

"(...) this first package of directives on maritime safety is **without doubt** one of the most important actions our House is going to implement in the course of 2000."

Indefinites

With the indefinites, we also find a lot of translations ending up in the 'residue'-category. For the other nouns, we find that again the majority of these stay in the same category. However, there are also ten indefinites in English that appear bare in Dutch. Dutch thus seems to be able to choose more freely whether or not to use a determiner after an \exists -based bare PP. Two examples of this behavior can be found below in (5.35) and (5.36).

(5.35) *En dat is wat ik bedoel als ik zeg dat we te maken hebben met een rechtsorde zonder beleid.*
 And that is what I mean when I say that we to do have with a law without policy.

"This is what I mean when I talk about a law **without a policy**."

- (5.36) *Ik denk dat dit een van de strijdpunten wordt die wij
I believe that this one of the bones of contention becomes which we
hopelijk met of zonder conferentie tot een goed eind kunnen brengen.
hopefully with or without conference to a good conclusion can bring.*
"I believe that this will become one of the bones of contention which we will hopefully -
with or **without a conference** - be able to bring to a successful conclusion."

Definites

From the frequency table, it is clear that definite NPs are quite uncommon, compared to bare forms and indefinite.¹³ For the cases that do exist, we see that half of them are also translated with the definite form in Dutch. However, in three occasions (25%!) we see the bare form arising:

- (5.37) (...) *landen zonder vetorecht zijn eerder deelstaten dan
(...) countries without right of veto are earlier constituent states than
zelfstandige naties.
independent nations.*
"(...) countries **without the right of veto** resemble constituent States more than
independent nations."
- (5.38) *Mijnheer de Voorzitter, er is zoals u weet een Engels gezegde dat luidt:
Mr the President, there is as you know an English says that says:
een show zonder ster is als een opvoering van Hamlet zonder eerste
a show without star is like a performance of Hamlet without first
doodgraver.
grave digger.*
"Mr President, there is as you know a saying in English that when the star does not appear
in a show it is rather like a performance of Hamlet **without the first grave digger.**"
- (5.39) (...) *[Ik kwam] even in de verleiding om te zeggen dat dit een Hamlet
(...) I came somewhat in the temptation to say that this a Hamlet
zonder prins zou gaan worden (...)
without prince would be (...)*
"(...) I was tempted to say that this would be like Hamlet **without the prince** (...)"

On a closer look, there are some hitches with these examples. First of all, it is questionable whether 'the right of veto' ('vetorecht' in Dutch) in (5.37) actually is a count noun. And even so, it is far more abstract than, for example, 'the prince' in (5.39). However, this latter example also has its problems. It seems here that 'the prince' is referring to a capacity (in the sense of [de Swart et al., 2007]). While capacity nouns in English must be preceded by a definite determiner, in Dutch, capacity nouns appear as bare nouns. The minimal pairs (5.40) and (5.41) below (from [Le Bruyn, 2010]) show this behavior.

(5.40) Jan is dokter.

(5.41) John is a doctor.

One could say that in the examples (5.38) and (5.39) above, a similar thing is happening: since Dutch is able to use capacity nouns bare in all contexts, speakers of Dutch can drop the determiner, while speakers of English have to use a definite determiner. But frankly, the conundrum here is not solved yet. First of all, as is clear from (5.41), English is obliged to use an indefinite to

¹³A hypothesis here is that this is because of a clash of the semantics of the preposition and the determiner. While 'without' selects for non-existing entities, 'the' refers to an existing, introduced entity. However, the data presented here is far too limited to draw such a conclusion.

refer to capacity (a definite here, 'John is the doctor', would change the semantics). However, in both examples with 'without' above, a definite is used. Secondly, English does give its speakers to leave the capacity bare when it refers to a unique position (as in (5.42) below, again copied directly from Le Bruyn's PhD thesis, show).

(5.42) Hyacinth is treasurer of the local Women's institute.

In Hamlet, the prince and (even more certainly) the first grave digger indeed are unique characters, and we thus would expect a drop of the determiner. But as said before, in the examples with 'without' we find a definite. Nevertheless, one can thus comment on the examples found in this corpus study. The earlier significant result found is therefore not totally beyond all doubt. However, a larger corpus study should be able to make more sense of the hypothesis that Dutch can more easily have bare nouns following an \exists -based bare PP.

5.3.3 Intermediate Conclusion

In this section on 'without' we have made an important observation: in comparison with English, Dutch more freely allows bare nouns. This becomes clear from Table 5.5 above. Here, we find that while the noun phrases often are translated with the same 'article'-category (bare, indefinite or definite), in English, one is far more likely to use a stronger, indefinite or definite alternative as a translation of bare forms in Dutch. We found that this result was statistically significant.

A closer look at the data, however, revealed that some of the data was somewhat flawed, and so we are yet unable to fully falsify our null hypothesis. However, I hope that the previous did give the right directions, and presume that a further research will indeed show that Dutch can choose more freely for 'weaker' determiners in the environment of \exists -based prepositions like 'with' and 'without'. To strengthen this particular stance, I will use a similar, but distinct method to accumulate more evidence. In the next sections, I will work out this second corpus experiment.

5.3.4 A New Method

In the multilingual research performed until now in this chapter, we again and again included all results in which we found a bare PP in one language. In the translation, we then often found that another construction was used. This was indeed sensible for finding out the exact semantics of 'per', However, for 'without', another approach might be more appropriate. I here think of a method in which we exclusively search for those occurrences which have a 'without'-kind of preposition in both languages. This is called a **parallel search**; because we search for occurrences of prepositions in both the source and the target text.

There are some drawbacks to this method. First of all, one sort of sacrifices completeness. In this method, one simply deletes all occurrences of 'without' and 'zonder' which are not directly translated. It might well be that we then throw away some interesting data points. Furthermore, it is not always the case that 'without' in one English sentence is equivocally mapped unto 'zonder' in the Dutch translation. Since the corpus is sentence aligned, and not word aligned, there is always the possibility that we stumble upon a sentence in which 'zonder' is actually *not* the translation of 'without' in the source text.

The here proposed procedure thus still requires some manual labor, although the amount is much less than with the method performed earlier in this section. A clear advantage thus is that we are able to scrutinize a larger part of the EuroParl corpus, which could help making our conclusions more statistically powerful. Again, our null hypothesis would be that English and Dutch are similar in their use of articles after \exists -based bare PPs.

5.3.5 Results

Doing a parallel search for 'without' and its Dutch counterpart 'zonder' in the whole 2000 proceedings of the European Parliament yields 1112 hits. Of these hits, I extracted 63 cases in which

'without' is followed by a bare singular count noun. For Dutch, I extracted 73 examples. For each of these examples, I labeled the translations as either bare, indefinite or definite (the familiar forms from the section above), plural ('zonder belemmering' ⇒ 'without obstacles' (literally: 'without obstacle')) or verb ('without hesitation' ⇒ 'zonder aarzelen' (literally: 'zonder aarzel-ing')). Again, a residue category is included for some remaining examples, not to be included in any of the categories mentioned.¹⁴ This procedure then yields the frequency table in 5.6 below.

Table 5.6: For 'without' and 'zonder' followed by a bare singular count noun, the translation was categorized as either 'bare', 'indefinite', 'definite', 'plural' or 'verb'. Cases in which the continuation after 'without' or 'zonder' were not in one of these categories were put into the 'residue' group.

category	# EN ⇒ NL	% EN ⇒ NL	# NL ⇒ EN	% NL ⇒ EN
bare ⇒ bare	40	63,49%	34	46,58%
bare ⇒ indef	7	11,11%	22	30,14%
bare ⇒ def	1	1,59%	5	6,85%
bare ⇒ plural	10	15,87%	5	6,85%
bare ⇒ verb	1	1,59%	6	8,22%
residue	4	6,35%	1	1,37%
total	63	100,00%	73	100,00%

From Table 5.6 we are able to draw similar conclusions to that of Table 5.5. Again, we find that Dutch will more often keep close to the bare form (63,49% of the cases, versus 46,58% for English), while English is more likely to use an indefinite when a bare form arises in Dutch (30,14% vs. 11,11%). We however also find that both languages use plurals instead of bare forms (Dutch here is even more likely to do so than English) or verbs (here, English again is in the majority). Of both of these categories, two examples are provided below.

- (5.43) *Wij zijn tot de slotsom gekomen dat het de voorkeur verdient dat wij*
 We have to the conclusion come that it the preference deserves that we
ten aanzien van dit verslag overgaan tot stemming zonder debat.
 with regard to this report proceed to vote without debate.
 "We have agreed that it would be better to proceed to the vote **without debating** the report."
- (5.44) *In dat geval moet er gemeenschappelijk worden opgetreden, zonder aarzel-*
 In that case must there commonly be acted, without doubt
ing of compromis.
 or compromise.
 "Such an event would demand common action, **without wavering or compromise.**"
- (5.45) *Dat is allemaal zonder trauma's verlopen, zonder moordpartijen, zonder*
 That is all without conflicts passed of, without murders, without
martelpraktijken of geweldpleging.
 acts of torture or acts of violence.
 "This was achieved without conflict, without murder, **without torture and without acts of violence.**"
- (5.46) *Het congres van de Tunesian Human Rights League [is] zonder incidenten*
 The congress of the Tunisian Human Rights League is without incidents
verlopen.
 passed off.
 "The congress of the Tunisian Human Rights League passed off **without incident.**"

¹⁴A particularly common one is the idiom 'zonder meer' as a translation of 'without doubt'.

Let us first of all argue about the use of the plural. The significant percentage of nouns that are translated with a plural instead of a singular, even with an effect of translation (we would expect that translators try to keep as close to the translation as possible, and thus, the amount of plurals would be expected minimal), suggests that bare singular forms after ‘zonder’ and ‘without’ can be substituted *salva veritate* by plurals.

A closer look at the semantics is corroborating evidence for this hypothesis. Most of the examples with ‘without’ or ‘zonder’ in this thesis, using either the singular or the plural form does not in any way change the meaning of the expression, as I have argued earlier in chapter 2 of this thesis. For example: if one has a house ‘zonder garage’ (‘without a garage’), one might as well say that one has a house ‘zonder garages’ (‘without garages’): the amount of garages stays zero. I would argue that $[[\textit{zonder} + N_{sg}]] \Leftrightarrow [[\textit{zonder} + N_{pl}]]$. For ‘with’ and ‘met’, this equality does not hold. I would thus expect that translators can not use a plural when translating from ‘with/met + N_{sg} ’. However, the data in this thesis seems not enough to either verify or falsify this hypothesis.

Next, we saw in the data that English is more likely to use a verb (more specific: a present continuous) as a translation of the bare singular in Dutch. The easy explanation for this behavior would be that Dutch lacks a present continuous, and therefore we did not find any of this behavior in Dutch. However, I would argue that there is another reason for English to use the verb form instead of a bare singular, indefinite or definite.

The argument runs as follows: in both Table 5.5 and 5.6, we saw that English was more likely to use an indefinite as a translation of a Dutch bare singular. However, this is not without a cost: using an indefinite introduces a discourse referent. Since Germanic languages according to [de Swart and Zwarts, 2009b] “avoid structure in the nominal domain”, using an indefinite is sort of a ‘last resort’: if there is a possibility not to introduce a discourse referent, English speakers ought to do so. With the present continuous, the English has (at least for the nouns that are nominalizations of verbs) an alternative at hand, in which no discourse referent is introduced.

5.4 Conclusion

In this section, we have compared \exists -based bare PPs in English and Dutch. By comparing the determiners (or lack thereof) after ‘without’ and its Dutch counterpart ‘zonder’, we were able to attest the null hypothesis that \exists -based bare PPs behave the same in both (closely related) languages. However, from two distinct methods we gained evidence that this null hypothesis can be falsified; and we should conclude that Dutch is more likely to follow their counterpart of ‘without’ by a bare singular count noun than English.

Chapter 6

Multilingual Corpora: Advantages and Technical Aspects

6.1 Introduction

In the last two chapters, we have seen the power of the multilingual corpora. And while there are yet quite a few multilingual corpora available (in this thesis, we have, for example, come across the EuroParl Corpus), researchers in both syntax and semantics just do not seem to use them very frequently. That might have three reasons. First of all, it might be that multilingual corpora are found to be hard to analyze: the methodology has yet to be developed. As a second problem, multilingual corpora yet just do not compare with monolingual corpora on size and annotating (e.g. POS-tagging). Especially this latter point seems to be quite a hurdle for effective corpus analysis. Last, but I expect certainly not least, researchers do not fully understand the concept of a multilingual corpus, and are therefore hesitated to use one for their research.

Let us one by one go over these reasons. Concerning the methodological questions, I hope to have given some roads to follow in the previous two chapters. With an effective explanation of what a multilingual corpus is all about, I hope to solve the latter problem somewhat. When multilingual corpus analysis becomes more of the norm in syntax and semantics, I am sure that soon the corpora will acquire the same status of monolingual corpora, solving also the second and last problem.

This chapter will proceed as follows. First of all, I will try to persuade readers that analysis with the help of multilingual corpora has several advantages compared to monolingual corpora. I will here also spend some time on problems with so-called *translation effects*, and show how to avoid these. After that, I will turn to one especially interesting technical detail with multilingual corpora: alignment. I will go over some important results on both sentence and word alignment.

In this chapter, I will use examples from a corpus I compiled myself, which consists of pairwise Dutch-English translations of "The Diary of A Young Girl" by Anne Frank.¹ This multilingual corpus is small in size (about 80.000 words), but tagged for parts of speech, and since literary multilingual corpora are hard to be found, might certainly be of interest to researchers of all kinds.

6.2 Advantages to Comparable Corpora

A devil's advocate could state that indeed, multilingual corpora are nice, but why not opt for several monolingual corpora that consist of a similar set of texts (e.g. newspaper articles). In [Johansson, 2007], some arguments for and against such comparable corpora are given. With such corpora, again, we are able to compare statistics for both languages, as we did in chapter

¹For Dutch readers, Anne Frank's diary is more commonly known as "Het Achterhuis".

2 and 3 of this thesis. A clear advantage would be that we do not have any problem of actually gathering the texts, as would be the case with multilingual corpora.

Furthermore, comparable corpora do not suffer from so-called source language influence. With translations of sentences, there is always the possibility that the author is more inclined to use the construction apparent in the source language, even though this construction is archaic or somewhat ungrammatical in the target language. In comparable corpora, all texts have been written in the mother's language, and thus there are no translation effects to be found.

However, there are at least two points against this line of reasoning. One that Johansson does not mention, is that it is not totally clear whether we do not find these translation effects in monolingual corpora. For example, with newspapers, at least some articles are translated from press releases written in another language than that of the paper. There could thus be some hidden translation effects. In multilingual corpora, the source language is often mentioned in the meta-text, and thus at least one knows where the influence come from. And furthermore, as Johansson rightly notes, if we are aware of the source language of the texts, we could opt to add some discounting factor to the statistics. In this manner we could control for translation effects.

Concerning the first argument provided above, that texts are not easy to be found, multilingual corpora do have the advantage that the texts that are chosen often have a certain value. That is, not every text that is written is translated, often only texts which have proven to be important (and the Diary of Anne Frank is of course a primary example of that). Because of this, these texts often are somewhat 'better', since multiple editors went over the text, clearing it hopefully of most spelling and grammatical errors. This could potentially mean that the data is more clean.

However, the translation of the texts might then be somewhat individual: it might be that another translator chooses just somewhat different words to express a similar meaning (an extreme case would be the well-known 'Jabberwocky' verse of Lewis Carroll from Alice in Wonderland). But we could argue that this 'individuality' of translation is likewise in other language use: then we also have our own ways to express some meanings. I thus do not think this is a good argument against multilingual corpora.

What on the other hand forms a big plus for multilingual corpora is its potential to not only provide cross-linguistic syntactic evidence: since the translated text should convey the same meaning and same discourse functions as in the source text, we are able to perform semantical analysis. In this thesis, we did so for 'per'. Of course, one could again say that there might be some translation effects, but it is clear that while comparable corpora do give us the option of syntactical comparison, semantical comparison seems to be a lot harder there.

All in all, I am obliged to conclude that the advantages of multilingual corpora outweigh the disadvantages brought up in this section. There is, however, one technical problem with multilingual corpora: how to assure which piece in the one text should be mapped upon a piece in the other. This problem of *alignment* will thus be the subject of the following section.

6.3 Alignment

An important next step in getting corpora ready for action is to align them. An alignment of two texts is, in fact, a mapping function: units of text of the source should be mapped upon units of text of the translation. This unit of text has traditionally been the sentence. However, current approaches, originally set up to provide better machine translation, also try to get a grip on word alignment.

6.3.1 Sentence Alignment

When we take the sentence as our unit, an alignment of two texts is the mapping function which tells us exactly which sentences from both the source and the translated text infer the same meaning. Most of the time, sentence boundaries are preserved across languages, and sentences are thus mapped one-to-one. However, when one translates a text, one often finds that in the one language, indeed, a punctuation mark is needed, while another language would prefer to keep

the sentence going. Some examples from the Anne Frank Corpus (all from the diary entries of January 1944) are provided below.

- (6.1) *Ik vind het best, dat ik hier een klein beetje mensenkennis gekregen heb,*
I find it fine, that I here a little bit human knowledge gained have,
maar het lijkt me nu voldoende.
but it seems me now enough.

"I've gained some insight into human nature since I came here, which is good, but I've had enough for the present."

- (6.2) *Maar Peter sprak doodgewoon verder over het anders zo penibele*
But Peter talked dead normal on about it otherwise very awkward
onderwerp, had helemaal geen nare bijbedoelingen en stelde me ten slotte in
subject, had at all no ulterior motives and put me eventually in
zoverre gerust, dat ik ook gewoon werd.
so much comfort, that I also normal became.

"But Peter went on talking in a normal voice about what is otherwise a very awkward subject. Nor did he have any ulterior motives. By the time he'd finished, I felt so much at ease that I started acting normally too."

- (6.3) *Een tijd lang waren we veel samen, maar overigens bleef mijn liefde*
A long time were we a lot together, but aside stayed my love
onbeantwoord. Toen kwam Peter op mijn weg en ik kreeg een echte
unrequited. Then came Peter on my way and I got a real
kinderverliefdheid te pakken.
crush to get.

"For a long time we went everywhere together, but aside from that, my love was unrequited until Peter crossed my path. I had an out-and-out crush on him."

What we see here, is that in every example, the mapping is different. With the first example in (6.1), of course, the mapping is 1-to-1 (as is the case with most other sentences), in (6.2) and (6.3), the mapping is respectively 1-to-3 and 2-to-2. Theoretically, every N-to-N-mapping thus seems possible. This makes alignment quite a hard problem.

The problem can be solved either manually or automatically. While manual alignment of course suffices for smaller texts (and thus was done for the Anne Frank Corpus), for larger texts, automatic alignment seems unavoidable. The most prominent and insightful article on this topic was provided by [Gale and Church, 1993]. They created a program for automatic alignment, which works on the principle that equivalent sentences should roughly correspond in length; that is, longer sentences in the source language should correspond to longer sentences in the target language. Interestingly, not the amount of words, but rather on the amount of characters proved a better predictor of the length of the sentence.²

The success rate of Gale and Church's automatic sentence alignment comes close to 96%. While this percentage seems quite high on the surface, such a result means that one in every twenty-five sentences is mapped to a wrong translation. This could seriously affect the statistics made on a corpus later on, and there is thus some need for more control within the process. Church and Gale call for lexical constraints to improve the alignment. Two examples: when you know that 'house' is always translated with 'huis' in Dutch, one might expect that when the source uses 'house', the target text on which this particular part should map on, at least contains 'huis'. A similar thing can be said about dates: if one sentence contains '1943', the sentence to be mapped upon should also contain this year. This latter lexical constraint would in the setting of a diary certainly improve the alignment.

²This is only the case with languages that are alike in the available characters; a comparison between e.g. English and Chinese on number of characters in a sentence would most probably fail.

Another clear problem, especially found within the Anne Frank Corpus (because the revised English translation was found to include some passages the original Dutch version lacked), is when a sentence or paragraph in the one language is not found in the other. Gale and Church report that their alignment script is not able to deal with such ‘mappings’, and I am not aware of any approach that can.

But even when a sentence alignment has been done perfectly, finding relevant results still needs a lot of manual labor, as was clear from chapter 4 (where we failed to do statistical analysis on the EuroParl corpus) and 5 (where we succeeded in analyzing ‘per’ and ‘without’) of this thesis. What opportunities do we have to further help smoothen the analysis of multilingual corpora? Of course, there is the possibility of POS-tagging, but *word alignment* seems to have an even better perspective.

6.3.2 Word Alignment

Word alignment is the natural language processing task of identifying translation relationships among the words (or more rarely, we find ‘word’ alignment for multiword expressions, for example [Moirón and Tiedemann, 2006], who analyze Dutch multiword expressions) of a source and target language. Such an identification results in an annotated corpus in which for each word in the source language, the translation is marked, and vice versa.

Again, word alignment could be done either manually or automatically. However, manual word alignment would be a very time-consuming task, as the alignment has to be done on a word-by-word basis. Automatic alignment thus here seems to be the most likely choice. For automatic alignment, one can then choose for either a *unsupervised* or a *supervised* method. In the former, the alignment program is not given any information but the two (already sentence-aligned) texts, and will have to map words on the basis of appearance in both the source and the target sentence. This is, of course, not an easy task, but the brute force of a computer often helps to create sensible data. However, one might question the linguistic validity of such models.

In the latter approach, supervised learning, the program takes a usually small number of manually word-aligned sentences as its input, and from these, starts to make up its model. Here, one might also opt to include dictionary information, or even information about the syntactical structure (POS-tags, for example). These latter additions of the model add to the linguistic validity, and according to recent results, also to the quality of the alignment.

6.3.3 Conclusion

The conclusion of this chapter would be that indeed, multilingual corpora form the logical next tool in the toolbox of researchers in both syntax and semantics. Multilingual corpora have several advantages to (comparable) monolingual corpora, the most important one being to be able to compare the semantics of expressions. However, when using multilingual corpora, one should be aware of the possible translation effects that could have occurred during the process of translation.

Compiling a multilingual corpus, on the other hand, is a hard job. After the ‘solving’ of Gale and Church the problem of sentence alignment, recent research in the domain of statistical machine translation has focused on word alignment. While the first results are promising, it is clear that there is still a long way to go before words will be mapped almost perfectly upon its translation. However, when this mapping will succeed, I think it is safe to say that multilingual corpora will form a new Rosetta Stone.

Chapter 7

Concluding Remarks

In this concluding chapter, first of all, the results of this thesis will be summarized. In the second section, some suggestions for further research within the domain of bare PPs.

7.1 Summary of the Results

In the first two chapters of this thesis, we considered bare PPs in English and Dutch. We did so using the framework of [Le Bruyn et al., 2009]. These authors distinguished between three¹ kinds of bare PPs: P-based, N-based and \exists -based bare PPs. For each of these PPs, there is another factor that licenses the bare occurrence of the noun. For P-based bare PPs, the preposition opens a specific lexical slot for nouns of a specific category, as in (7.1) below. With N-based bare PPs, we see that nouns can appear bare with a multitude of prepositions, as in (7.2). It thus seems that the noun licenses the bare category. In the last category, exemplified in (7.3) below, we find \exists -based bare PPs. We here again see that the preposition ('with' or 'without') opens a slot, but this time for (possibly) all nouns.

(7.1) under N_{action} : control, consideration, discussion, study, review, arrest, construction, investigation, scrutiny...

(7.2) P school: about, after, at, during, from, in, through, to...

(7.3) without N: exception, analysis, comment, question, accusation, argument...

When we compared English and Dutch, we found that while English has more types of N-based bare PPs, Dutch has more \exists -based bare PPs. With regard to P-based bare PPs, we found that Dutch and English have a quite similar distribution. The results can be nicely summarized in Table 7.1, earlier presented in this thesis as Table 7.1.

The typology we set up for English and Dutch left some questions open. We tried to solve these using a multilingual corpus. In chapter 6, we went over the advantages, disadvantages and technical problems of such corpora, especially alignment. In chapter 4, we tried for a statistical analysis of the multilingual corpus. However, since the corpus we used (the EuroParl corpus) was not tagged for parts of speech, the research here was not as successful as one might have expected. However, the corpus was applicable for analysis per preposition, a road we took in the fifth chapter of this thesis.

Here, we found that the UNIT reading with 'per' was found far more frequently in Dutch, and that English employs all kinds of alternatives for this construction. Especially interesting was the use of the quantifier 'each', which is found to be strongly individuating. For 'per', we would then expect a similar semantics.

¹Actually four, but since D-based bare PPs are only to be found in Romanian, this fourth kind was not considered.

Table 7.1: A typology for bare PPs in English and Dutch. An asterisk (*) denotes that this form of bare PPs is far less used compared to the other language. A hyphen denotes the bare PP is absent in this language. The 'count' categories (many, some and zero) are only indicative, and should not be interpreted with any real number.

P-base bare PPs	Category of nouns	English preposition	Dutch preposition
	N ⇒ UNIT	per	per
	MODES OF TRANSPORT	by/on	per/te
	COMMUNICATION MEDIA	by	per
	RECORDING MEDIA	on	op
	MEDIA AS BUSINESS	in	-
	CLOTHING	in	in
	ACTION	under	onder*
	ACTIVITY	on*	op
N-based bare PPs	Category of nouns	# in English	# in Dutch
	LOCATION	many	some
	INTERRUPTION	many	zero
∃-based bare PPs	Preposition	# in English	# in Dutch
	<i>with</i>	zero	some
	<i>without</i>	some	many

In a second corpus research we focused on 'without'. In monolingual analysis, it seemed that English was less likely to have bare nouns after 'without'. Multilingual analysis confirmed this hypothesis: where Dutch used the bare form, English quite frequently used either an indefinite, definite or (when possible) a present continuous. Such a finding adds to the belief that articles in English and Dutch, while similar on the surface, enjoy quite a different status.

7.2 Suggestions for Further Research

Although in this thesis, we have been able to get to quite some results in the domain of bare PPs, there is - as always - more to do. First of all, the logical thing to do would be to fill in the typology of bare PPs, given in 7.1, for other languages. In that way, we would get a clearer image of the bare PP spectrum across languages.

Another logical extension of this thesis would be to do a comparison of N-based bare PPs in the multilingual corpus. For the results presented in the last paragraph, we used comparable corpora, and it would be nice to get additional evidence for the points made by using multilingual corpora. Then, we might also be able to verify the semantic conclusions that were drawn in e.g. [Stvan, 2009]. Of course, the multilingual analysis given for 'per' could also be extended to other P-based bare PPs.

A domain which have largely left aside in this thesis is that of modification of bare PPs. In the introduction of this thesis, I mentioned that modification of bare PPs usually lead to ungrammaticality, but there are some clear exceptions to this rule, as the example in (1.6) should show, copied below in (7.4). This domain of restricted modification could be given a clearer overview by some methods exploited by [van der Beek, 2005] in Dutch. In her PhD-thesis, she also did some research on obligatory modification: some nouns need a modifier to be able to appear without determiner at all. Some examples of this latter behavior (one in Dutch, one in English) are copied in (7.5) and (7.6) below.

(7.4) Pat is in *big / *red / federal / state prison.

(7.5) op vegetarische/politieke/water-/... basis (on vegetarian/political/water/... basis)

(7.6) at great/public/considerable expense

Of course, there are many more roads to travel, but it seems the road map for research into bare PPs now is quite clear, as we can start from the given typology, and extend our analysis further when needed. From this thesis, I hope it became obvious that insight into bare PPs can come from either a monolingual or - sometimes preferably - from a multilingual perspective.

Bibliography

- [Baldwin et al., 2006] Baldwin, T., Beavers, J., van der Beek, L., Bond, F., Flickinger, D., and Sag, I. A. (2006). In search of a systematic treatment of determinerless PPs. In Saint-Dizier, P., editor, *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*, pages 163–179. Kluwer Academic Publishers.
- [Bird et al., 2009] Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O’Reilly Media Inc.
- [Blutner, 2000] Blutner, R. (2000). Some Aspects of Optimality in Natural Language Interpretation. *Journal of Semantics*, 17:189–216.
- [Borthen, 2003] Borthen, K. (2003). *Norwegian Bare Singulars*. PhD thesis, Norwegian University of Science and Technology.
- [Chierchia, 1998] Chierchia, G. (1998). Reference to Kinds Across Languages. *Natural Language Semantics*, 6:339–405.
- [Davies, 2004] Davies, M. (2004). BYU-BNC: The British National Corpus. Available at: <http://corpus.byu.edu/bnc>.
- [de Swart et al., 2007] de Swart, H., Winter, Y., and Zwarts, J. (2007). Bare nominals and reference to capacities. *Natural Language and Linguistic Theory*, 25(1):195–222.
- [de Swart and Zwarts, 2009a] de Swart, H. and Zwarts, J. (2009a). Less form–more meaning: Why bare singular nouns are special. *Lingua*, 119(2):280–295.
- [de Swart and Zwarts, 2009b] de Swart, H. and Zwarts, J. (2009b). Nominals with and without an article. In Hendriks, P., de Hoop, H., Kramer, I., de Swart, H., and Zwarts, J., editors, *Conflicts in Interpretation*, pages 109–136. Equinox, London.
- [Dömges et al., 2007] Dömges, F., Kiss, T., Müller, A., and Roch, C. (2007). Measuring the productivity of determinerless PPs. In Costello, F., Kelleher, J., and Volk, M., editors, *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 31–37. Omnipress.
- [Farkas and de Swart, 2003] Farkas, D. and de Swart, H. (2003). *The semantics of incorporation: from argument structure to discourse transparency*. Stanford: CSLI Publications.
- [Fodor and Pylyshyn, 1988] Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critique. *Cognition*, 28:3–71.
- [Gale and Church, 1993] Gale, W. A. and Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102.
- [Haspelmath, 1997] Haspelmath, M. (1997). *From space to time: Temporal adverbials in the world’s languages*. Newcastle: Lincom Europa.
- [Jackendoff, 2008] Jackendoff, R. (2008). Construction after construction and its theoretical challenges. *Language*, 84(1):8–28.

- [Johansson, 2007] Johansson, S. (2007). *Seeing through multilingual corpora: on the use of corpora in contrastive studies*. John Benjamins Publishing Company.
- [Koehn, 2005] Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*. MT Summit.
- [Kucera and Francis, 1967] Kucera, H. and Francis, W. (1967). *Computational analysis of present-day American English*. Brown University Press.
- [Le Bruyn, 2010] Le Bruyn, B. (2010). *Indefinite articles and beyond*. PhD thesis, Universiteit Utrecht.
- [Le Bruyn et al., 2009] Le Bruyn, B., de Swart, H., and Zwarts, J. (2009). Bare PPs across languages. Workshop on Bare nouns: Syntactic projections and their interpretation.
- [Longobardi, 1994] Longobardi, G. (1994). Reference and Proper Names: A Theory of N-Movement in Syntax and Logical Form. *Linguistic Inquiry*, 25(4):609–665.
- [Mardale, 2006] Mardale, A. (2006). Why *on table* is *on the table*? In Gyuris, B., editor, *Proceedings of the First Central European Student Conference in Linguistics*, pages 109–136. Research Institute for Linguistics of the Hungarian Academy of Sciences, Budapest.
- [Moirón and Tiedemann, 2006] Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In Rayson, P., Sharoff, S., and Adolphs, S., editors, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40. East Stroudsburg, PA: Association for Computational Linguistics.
- [Paenen, 2009] Paenen, M. (2009). *PP's met kale nomina in het Nederlands*. BA Thesis, Universiteit Utrecht.
- [Pustejovsky, 1991] Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.
- [Stvan, 1998] Stvan, L. (1998). *The semantics and pragmatics of bare singular noun phrases*. PhD thesis, Northwestern University.
- [Stvan, 2007] Stvan, L. (2007). The Functional Range of Bare Singular Count Nouns in English. In Stark, E., Leiss, E., and Abraham, W., editors, *Nominal Determination: Typology, Context Constraints, and Historical Emergence*, pages 171–187. John Benjamins, Amsterdam.
- [Stvan, 2009] Stvan, L. (2009). Semantic incorporation as an account for some bare singular count noun uses in English. *Lingua*, 119(2):314–333.
- [Tunstall, 1998] Tunstall, S. L. (1998). *The Interpretation of Quantifiers: Semantics and Processing*. PhD thesis, University of Massachusetts.
- [van der Beek, 2005] van der Beek, L. (2005). *Topics in Corpus-Based Dutch Syntax*. PhD thesis, Rijksuniversiteit Groningen.
- [van Grootheest, 1989] van Grootheest, D. (1989). *Eindhoven Corpus (VU-versie)*. Available at: <http://www.inl.nl/>.
- [Zwarts and Winter, 2000] 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.