Chapter 5

The empirical base of semantic research

5.1 Introduction

In chapters 2 and 3, we sketched an approach to genericity based on Gerstner-Link & Krifka (1993), Krifka et al. (1995), Link (1995), Carlson (1995), Chierchia (1995, 1998), De Swart (1996), Cohen (1999, 2001), Longobardi (2001), Rimell (2004) and other works in the same framework. Most of the judgements on syntactic and semantic well-formedness in these works are based on the researcher’s own intuitions. A limitation of research based on the intuitions of one speaker is that it remains unclear how much variation there is among speakers of the same language. A researcher who claims that his results are representative of Dutch (or another language) in general needs a more extensive empirical base. Investigations of the judgements of groups of informants and corpus-based studies enable the researcher to draw more representative conclusions.

This chapter discusses limitations and advantages of intuition-based and corpus-based approaches. The chapter is organized as follows. Section 5.2 discusses the limitations and advantages of intuition-based research. We will discuss research based on the intuitions of one speaker as well as research based on the judgements of a number of native speakers of the language(s) or dialect(s) under consideration. Section 5.3 is devoted to a discussion of the merits and demerits of corpus-based approaches. The conclusions are summarized in 5.4.
5.2 Intuition-based approaches

5.2.1 Introduction

This section discusses the use of acceptability judgements as a tool for linguistic research.\textsuperscript{1} The acceptability judgements in this study are based on two kinds of intuition-based methods:

- questionnaire-based research into the acceptability judgements of speakers of a number of local (or regional) varieties\textsuperscript{2} of Dutch.
- investigations of the acceptability judgements of small groups of native speakers of Standard Dutch (i.e. groups of between 5 and 10 speakers).

In section 5.2.2, we examine some distinctions between different types of intuitive judgements. Section 5.2.3 discusses some advantages and provides some justification of the use of intuitions. Section 5.2.4 is devoted to the limitations of intuition-based approaches. Section 5.2.5 summarizes how intuition-based approaches are used in this thesis.

5.2.2 Types of intuitions

Wasow & Arnold (2005: 1482 and 1489) make a distinction between ‘primary’ and ‘secondary intuitions’:

“Two types of intuitions have played a central role in linguistic research over the past half century. The first, which we will call ‘primary intuitions’, are simply introspective judgements of a given linguistic expression’s well-formedness or of its meaning. The second, which we dub ‘secondary intuitions’, are intuitions about why a given expression is (or is not) well-
formed or has the meaning it has. (...) it is an intuition about how to analyze the unacceptability of a given sentence, not about the acceptability itself.”

Wasow & Arnold (2005: 1488) point out that “investigators in every discipline have intuitions about what constitutes a plausible explanation”, but that these intuitions do not “constitute empirical evidence, and their role should be subordinate to primary data”. A researcher should not assign a more important role to secondary intuitions. Secondary intuitions will not be discussed further in this chapter.

Primary intuitions can be classified along the following lines:

- one can judge sentences to be unacceptable under any interpretation or under some interpretation(s) (cf. 5.2.2.1).
- well-formedness intuitions can be categorical or gradient (cf. 5.2.2.2).

The following sections discuss some implications of the use of different kinds of intuitions.

5.2.2.1 Acceptability judgements about sentences under certain interpretations

Sentences can be judged (un)acceptable from a syntactic perspective, but they can be semantically (un)acceptable as well. In the latter case, the acceptability often depends on the interpretation assigned to the sentence. Noël (2003: 8) discusses a potential problem of such acceptability judgements (cf. also Colleman 2006: 195):

“(...) treatments of clausal complementation (...) have maintained that different kinds of complement always convey different meanings [i.e. different kinds of complement are well-formed under different meanings, AO]. These meanings are usually illustrated with a few well-chosen construed and decontextualized sentences, accompanied by a couple of starred sentences, i.e. by claims that the meanings of these complements preclude certain matrix verb-complement combinations. Readers are then expected to match these claims about meanings and predictions about what are and what are not possible sentences against their own intuitions and if there is no ensuing disagreement these claims and predictions are taken to be proven. That is to say, they are taken to be proven without having been checked against any kind of linguistic data (other than a couple of individuals’ intuitions (...)) (...) Probably no other scientific discipline would take the kind of argumentation presented in these studies seriously (...). Grammaticality judgements (...) can be very idiosyncratic (...).”

It is an accurate observation that grammaticality/acceptability judgements are idiosyncratic. However, the same problem arises if one wishes to investigate correspondences between form
and meaning by carrying out corpus studies. In such corpus studies, the corpus researcher has to determine which interpretations the sentences in his corpus (can) receive. This means that the reader has to trust the primary intuitions of the researcher. This claim will be illustrated with an example in 5.3.3.1.1.

5.2.2.2 Gradient well-formedness

Speakers sometimes judge sentences neither acceptable nor unacceptable, but somewhere in between the two. Schütze (1996: 41-48) discusses such subtle judgements. He cites Belletti & Rizzi’s (1988) influential work on psych-verbs in Italian and reports that these authors distinguish between “no fewer than seven degrees of grammaticality”.

An example of a subtle judgement that is relevant to this study can be found in Broekhuis, Keizer & Den Dikken (2003: 609). These authors prefix sentence (1) with an asterisk and a question mark. This implies that the sentence is relatively ungrammatical. In section 5.2.4.1, I will show that other resources consider sentences such as (1) to be well-formed. So, there is inter-speaker variation, which is indicated by the symbol ‘%’. In cases of gradient well-formedness and inter-speaker variation, it is especially important to gather the intuitions of a number of native speakers (unless we are exclusively interested in the I-language of one individual).

(1) % Dodo’s zijn uitgestorven.
   ‘Dodos are extinct.’

5.2.3 Justification of the use of the researcher’s intuitions

This section provides some justification of research in which acceptability judgements are based on the researcher’s intuitions. In the introduction to this chapter, we have observed that (most of) the acceptability judgements in the (formal) semantic publications under discussion in this study are based on the researcher’s intuitions. This procedure is not unique for (formal) semanticists. Syntacticians working in the generative tradition in general follow the same methodological procedures. Borsley & Ingham (2002: 1) refer to the frameworks that follow this approach as “mainstream theoretical linguistics”. They write:

“Mainstream work is mainly concerned with I-language, the cognitive system underlying the ordinary use of language. Various applied linguists have suggested that this is unsatisfactory and that texts are a more appropriate object of study. (…) Kress (1993) (…) is under the
impression that mainstream linguists believe that all members of a speech community have the
same I-language and use it in the same way, but of course no one believes this. (…) A second
important feature of mainstream linguistics is a reliance on native speakers’ intuitions as a
major source of data. Various applied linguists think this is unsatisfactory, and suggest that
corpora of naturally occurring speech and writing are a more satisfactory source of data.”

(Borsley & Ingham 2002: 2-3)

The emphasis on I-language, the mentally represented linguistic knowledge that a native
speaker of a language has (see Chomsky 1988), implicitly underlies many theoretical
syntactical and semantic studies. In some cases, this perspective is made explicit, as in
Haegeman (2004: 120):

“Uiteraard is niet uitgesloten dat andere sprekers (...) andere oordelen hebben, maar dit is op
zich te verwachten vanuit de generatieve visie op de grammatica met haar nadruk op de I-
language, de interne taal, d.w.z. de competentie van de individuele taalgebruiker.”

(Of course it is not impossible that other speakers make other judgements. This is expected
from the generative view on grammar, with its emphasis on the I-language, the internal
language, i.e. the competence of the individual language user.)

Under this perspective, making an acceptability judgement does not necessarily imply that all
speakers of the relevant language share it. The ultimate consequence of using the intuitions of
one speaker as a tool for linguistic research is that the data are representative only of one
idiolect. So, if we wish to investigate the internal consistence in a system of judgements
corresponding to the I-language of one speaker, the best thing we can do is using intuition-
based methods.

A second point has to do with the subjectivity of intuition-based results. Schütze
(1996: 1) writes:

“I argue that if several simple steps are taken to remove obvious sources of bias,
grammaticality judgments can provide an excellent source of information about people’s
grammars. Thus, I respond to two of the most widespread criticisms of generative grammar –
namely, that it involves constructing theories of intuition rather than of language use, and that
it is highly subjective and biased by the view of the linguist.”

(my underlining, AO)

Subjectivity is (or can be) a welcome aspect of intuition-based results. Corpus results (often)
do not tell us anything about inter-speaker variation. Corpus results do not provide an
appropriate basis for drawing conclusions about how many (and what) speakers share a

3 “I-language” and “competence” can be considered as two terms for the same concept (cf. Adger 2003: 17).
judgement. A more appropriate method to describe this variation is by using the subjective intuitions of speakers of a language. Such research can be carried out in a systematic fashion by using questionnaires and gathering the intuitions of a number of native speakers (cf. 5.2.5).

A third point is that intuition-based claims are verifiable to a large extent. This point can be illustrated by sentence (2a). Zwart (1997: 28) claims that the use of vandaag in this position is well-formed. Zwart writes that “[t]opicalizations (...) invariably trigger inversion of the subject and the verb in tensed main clauses. The topic(...)-element and the finite verb are strictly adjacent. The finite verb and the subject no longer have to be adjacent” (my underlining, AO). The observation that the topic-element and the finite verb are strictly adjacent is illustrated in (2b). The claim that the finite verb and the subject “no longer have to be adjacent” is illustrated by sentence (2a). However, a number of speakers of Dutch judge sentence (2a), in which vandaag appears between the finite verb and the subject, to be unacceptable. This claim is based on the intuitions of nine native speakers of Dutch. Eight of these nine informants judge the sentence unacceptable. So, Zwart’s observation is not representative of Dutch in general.4 His observation does not apply to the I-language of speakers of Dutch who do not accept sentences like (2a).

(2) a. Marie kussen (%vandaag)de jongens vaak.
    marie kiss (today) the boys often
    ‘Mary the boys kiss a lot today.’

b. Marie (*vandaag) kussen de jongens vaak.
    marie (today) kiss the boys often

The crucial point to be made here is that the representativity of Zwart’s observation can be easily verified by consulting native speakers, as illustrated above. In many cases, corpus results are less easily verifiable. Van der Beek (2005), who applies corpus data in various kinds of linguistic analyses and shows that corpus data “is a very useful resource for syntactic research” (cf. abstract), writes:

“In the evaluation step, the good examples are extracted from the total set of candidates. (...) This set will (...) contain sentences which are grammatically fine, but do not contain the linguistic structure under investigation. (...) On top of that, the corpus is likely to contain a number [of] ungrammatical sentences. (...) One can conclude that there is still an important role for grammaticality judgments in a corpus-based approach: separating the true positives [i.e. grammatical sentences] from the false positives [i.e. ungrammatical sentences].”

(Van der Beek 2005: 3, my underlining, AO)

4 It could be that sentence (2a) is grammatical, even though it is unacceptable (cf. note 1). Possibly, the sentence can be assumed to be “hard to process”. However, Zwart does not present arguments in favour of such a claim.
So, in such corpus investigations, ill-formed sentences are filtered out from the results. Usually, it cannot be verified on the basis of what acceptability judgements sentences were filtered out (unless the study contains a complete list of which sentences are included and which ones are filtered out).

5.2.4 A limitation of intuition-based approaches: conflicting intuitions

In 5.2.3, we have argued that one of the advantages of intuition-based approaches has to do with the fact that acceptability judgements are subjective. An appropriate method to describe the variation in acceptability judgements is by using the subjective intuitions of native speakers of a language. This section shows that the subjectivity of acceptability judgements is not always an advantage, but can be an important limitation of intuition-based approaches.

In 5.2.4.1 and 5.2.4.2, I will demonstrate that four publications, namely Haeseryn et al. (1997), Devos, De Muynck & Van Herreweghe (1991), Broekhuis, Keizer & Den Dikken (1997) and Ter Meulen (1995) make conflicting grammaticality judgements about two types of sentences. This is a direct consequence of the use of intuition as a method in linguistic research. Conflicting intuitions make clear that the intuition-based approach is not a solid basis for drawing general conclusions about Dutch. A researcher who claims that his results are representative of Dutch (or another language) in general needs a more extensive empirical base.

5.2.4.1 Bare plurals in subject position of kind predicates

In the literature there is no consensus on the well-formedness of sentences like (3), in which a bare plural is used in subject position of a kind predicate. It is important to note that the well-formedness of sentences such as (3) has further-reaching implications than might be obvious at first glance. No one would assume that bare plurals refer to kinds in characterizing or episodic sentences if sentences like (3) were ill-formed (cf. the discussion in chapter 3). In varieties of Dutch that do not accept sentences like (3), bare plurals do not get kind-referential interpretations at all. As a consequence, there would be a one-to-one relation between definiteness and kind reference in such varieties: noun phrases with a definite article refer to kinds, while noun phrases introduced by an indefinite article receive non-kind-referring readings.
(3) a % Pinguïns sterven uit.
   Penguins die out
   ‘Penguins are dying out.’

b % Olifanten sterven uit.
   Elephants die out
   ‘Elephants are dying out.’

In (formal) semantic as well as in traditionally oriented, descriptive publications, the judgement has been made that sentences such as (3) are well-formed. In Devos, De Muynck & Van Herreweghe (1991: 43), sentence (3a) is not marked with an asterisk. These authors explicitly state that sentence (3a) expresses an assertion about “de soort” (the kind). From the description by Devos, De Muynck & Van Herreweghe, we can conclude that according to this resource, sentence (3a) is well-formed under a kind-referential reading. Ter Meulen (1995: 356) makes the same judgement.

The opposite intuition can be found as well. In the terminology adopted by Haeseryn et al. (1997) a noun phrase is ‘generic’ if it can be combined with a kind predicate (see chapter 4). Haeseryn et al. (1997: 812) assume that bare plurals do not receive the ‘generic’ interpretation. This implies that according to this grammar sentences like (3) are ill-formed. A comparable judgement can be found in Broekhuis, Keizer & Den Dikken (2003). They present an example sentence in which a kind predicate is combined with a bare plural. The relevant sentence is prefixed with an asterisk and a question mark (cf. Broekhuis, Keizer & Den Dikken (2003: 609), my emphasis, AO):

"The fact that definite noun phrases may refer to the whole class, whereas indefinite noun phrases cannot is clear from the examples in (29). The examples in (29b&c) are semantically anomalous since the predicate uitgestorven ‘extinct’ can be predicated of a species as a whole, but not of the individual members of a species.

(29) (...) b  * Een Dodo is uitgestorven. (...)
   c  *? Dodo’s zijn uitgestorven.”

In cases like this, the results of corpus studies can be relevant. Some relevant corpus results will be presented in section 5.3.2.1.

5.2.4.2 Definite plurals in characterizing sentences

In this section, I will present a second example of conflicting intuitions in the literature. In the literature there is no consensus on the well-formedness of sentences like (4), in which a (non-

The opposite intuition can be found as well. Haeseryn et al. and Broekhuis, Keizer & Den Dikken judge sentences like (4) ill-formed under the relevant reading. Sentence (4c) is taken from Haeseryn et al. (1997: 811) and is judged ill-formed by the authors.

Broekhuis, Keizer & Den Dikken (2003: 612-613) state that sentence (4d) can only be interpreted as a statement about “a contextually determined group of zebras”. According to this resource, example (5) is more acceptable, because grote kat (‘big cats’) can be interpreted as “the name of the superset containing the subsets of cats denoted by the nouns leeuw ‘lion’, tijger ‘tiger’, etc. In other words, the noun phrase de grote katten does not refer to one, but to several species of animals, hence its plural form” (cf. Broekhuis, Keizer & Den Dikken 2003: 612).
De grote katten zijn gevaarlijke roofdieren.
‘The big cats are dangerous predators.’

The relevant interpretation of (5) is available for its English equivalent as well (cf. the English translation). This interpretation is a taxonomic one. Krifka et al. (1995:74) observe that taxonomic noun phrases show the whole range of syntactic behaviour that we observe with every count noun (cf. section 3.5). Thus, we predict that definite plurals receive taxonomic interpretations, just like other common noun phrases. Sentence (5) makes clear that this prediction is borne out. This does not contradict the claim that Broekhuis, Keizer & Den Dikken judge sentences such as (4d) and (5) to be unacceptable under a non-taxonomic characterizing reading.

In section 5.3.2.2, I will present some results of corpus studies into the frequency of sentences such as (4). Corpus findings can be interpreted as indirect evidence on whether there are speakers who judge a type of sentence well-formed.

5.2.5 Intuition-based research in this thesis

5.2.5.1 Investigations of the judgements of groups of speakers

In the previous sections, we have discussed the merits and demerits of intuition-based results. An important point in that discussion was that intuitions of native speakers are subjective. This is an important limitation of research based on the intuitions of one speaker. An advantage of the subjectivity of acceptability judgements is that inter-speaker variation can be described by using the subjective intuitions of native speakers of a language. In 5.2.4, we have discussed two cases in which conflicting judgements about the well-formedness of the same sentence type are made in the literature. On the basis of such conflicting judgements, we can draw the conclusion that the relevant sentences are subject to inter-speaker variation.

In the two cases discussed in 5.2.4, the existing literature enables us to draw conclusions about inter-speaker variation. Usually, this type of evidence (i.e. conflicting intuitions in the literature) is not available. In such cases, we have to collect the acceptability judgements of a group of native speakers.

In general, the judgements presented in this study are not meant to represent an idiolect. If there is evidence that there is variation among speakers, this is indicated by using the symbol ‘%’. In many cases, claims about inter-speaker variation in the judgements about acceptability will be based on investigations of the judgements of small groups of native speakers. As a consequence of the fact that the results are based on small numbers of informants (i.e. groups of between 5 and 10 native speakers), we cannot draw conclusions
about what percentages of native speakers of Dutch do and do not accept the sentence types we are interested in. Our findings only indicate whether speakers agree that a sentence is acceptable, agree that the sentence is unacceptable or disagree about the acceptability of the sentence.

5.2.5.2 Questionnaire-based research

Questionnaires are an important tool for obtaining acceptability judgements from speakers of a language. In chapter 6, I will present the results of a questionnaire study. This study does not focus on the variation among speakers of Standard Dutch, but investigates acceptability judgements of speakers of local varieties of Dutch.

In our questionnaire study, language users were asked to rate the acceptability of a number of sentences on a scale from 1 to 5, where the value ‘1’ corresponds to ‘completely unacceptable’ and the value ‘5’ corresponds to ‘fully acceptable’. This scale has been used in many other recent syntactic and semantic studies (cf. Van Craenenbroeck 2004, Hoeksema & Den Ouden 2005 and Clifton, Fanselow & Frazier 2006, among many others).

Whether this method is the most appropriate and sensitive method to investigate acceptability judgements is a question which has generated some debate. Bard, Robertson & Sorace (1996: 35) write that there is a

“disproportion between the fineness of judgements people can make and (...) the five-point scale often used in empirical studies or indeed to any other scale that predetermines the number of distinctions (...). There is no way of knowing in advance if our sensitivities are limited to a five-way distinction any more than a four-way distinction.”

Bard, Robertson & Sorace describe how magnitude estimation, a technique used in psychophysics, can be adapted for eliciting acceptability judgements. Magnitude estimation, applied to linguistic acceptability requires the subject (or the experimenter) to associate a numerical judgement with a sentence. Once the initial sentence is presented and a number associated with it by the subject (or the experimenter), the subject assigns to each successive sentence a number reflecting the acceptability of the sentence in comparison with the first sentence. Subjects are instructed to reflect perceived ratios in their judgements: a sentence that appears to be 2 times as acceptable as the first is to be given a number 2 times the original number.

According to Bard, Robertson & Sorace (1996: 41), this technique has some important advantages over more traditional methods, like the five-point scale often used in empirical studies: “(...) it does not restrict the number of values which can be used (...). Subjects decide whether each stimulus should be assigned the same number as another stimulus or a different
number, and they have complete freedom about which of the infinite set of numbers to use.” I refer the reader to Alexopoulou & Keller (to appear) for an application of magnitude estimation.

Although the technique of magnitude estimation offers interesting opportunities for the (future) empirical study of (Dutch) generics, it is not applied in the questionnaire study presented in the following chapter. The results of our questionnaire study are based on the 5-point scale. The 5-point scale is the prevailing method in contemporary semantic and syntactic theory and has been (successfully) applied in the study of Dutch dialects (cf. Van Craenenbroeck 2004).

5.3 Corpus-based approaches

5.3.1 Introduction

There have been many discussions about the role of corpora in (theoretical) linguistics. Verkuyl (1998: 63) defends the Anti-Corpus-axiom:

“het nut van corpora voor het doen van taalkundig onderzoek is zo minimaal, dat het zonde van de tijd is om ze statistisch te gebruiken”

(the benefit of corpora for linguistic research is so minimal, that it is a waste of time to use them statistically)

In a later edition of the same journal (Nederlandse taalkunde), Baayen (1998) argues against this axiom and claims that careful and statistically valid corpus research is potentially relevant to many issues of theoretical linguistics (cf. Kruyt 1998 for another reaction). Other debates between introspective and corpus-oriented linguists can be found in Borsley & Ingham (2002, 2003) and Stubbs (2002) and in Newmeyer (2003, 2005), Clark (2005), Laury & Ono (2005) and Meyer & Tao (2005). In such debates, the use of corpora is contrasted with the use of linguists’ own introspective judgements as a resource for theoretical linguistic inquiry. Arguments against one of these two methods can often be reformulated as arguments in favour of the other method.

Chomsky’s (1957, 1965) main objection against corpus-based approaches is that according to him the linguist must seek to model language competence (or I-language) rather than performance. This brings us to the essence of Chomsky’s criticism: A corpus is a collection of externalized utterances; it provides performance data, and as such it is a poor
guide to investigating linguistic competence (cf. McEnery & Wilson 2001). How far-reaching the consequences of this argument are depends on one’s commitment to the Chomskyan view on the main task of the linguist. Section 5.3.3 sets out some demerits of corpus research from a more general perspective.

Section 5.3.2 is devoted to the advantages of corpus data. Section 5.3.3 discusses some drawbacks of corpus-based methods. Section 5.3.4 discusses how corpus-based approaches are used in this thesis.

5.3.2 Advantages of corpus-based results

In section 5.2.4, we have presented two examples of conflicting judgements made by linguists judging the well-formedness of the same sentence type. The relevant sentence types are kind predicate sentences with bare plural subjects (cf. (3), repeated as (6)) and characterizing sentences with definite plural subjects (cf. (4), repeated as (7)).

(6) a. % Pinguïns sterven uit.
b. % Olifanten sterven uit.

(7) a. % De olifanten hebben waardevolle tanden.
b. % De Amerikanen houden van glamour.
c. % De paarden zijn zoogdieren.
d. % De zebra’s zijn gestreept.

We have argued that conflicting judgements are unavoidable when using intuition as a tool for linguistic research. A more extensive empirical base is needed if a researcher claims that his results are representative of Dutch (or another language) in general. In 5.2, we have discussed that gathering the intuitions of a number of native speakers enables the researcher to draw more representative conclusions. In the following sections, I will illustrate that corpus-based approaches are a second method to obtain more representative conclusions about whether speakers of a language judge a certain sentence type well-formed. Suppose, for example, that bare plurals are used with some regularity in subject position of the verb *uitsterven* (‘die out’). Then we would have some evidence supporting the claim that sentences like (6) are acceptable at least for some speakers.
5.3.2.1 Bare plurals in subject position of kind predicates

This section presents the results of a corpus study into the frequency of definite singular, definite plural, indefinite singular and bare plural common noun phrases in subject position of *uitsterven*. The results are based on three corpora: the *INL 27 Miljoen Woorden Corpus 1995*, the *INL 38 Miljoen Woorden Corpus* and the *CONDIV*-corpus. These corpora contain a number of varieties of written Dutch (academic prose, juridical texts and other official documents, material from magazines, material from quality and popular newspapers, texts from non-fiction books, internet language from chat channels and discussion groups). The total corpus contains approximately 110 million words.

This corpus contains 60 sentences in which the verb *uitsterven* is used with a count noun phrase in subject position. Sentences in which *uitgestorven* can be analyzed as a predicatively used adjective are also included. All these sentences originate from texts published in the period 1970-2000. The results are presented in table 5.1.

Table 5.1: The frequency of noun phrase types in subject position of *uitsterven*/*uitgestorven*

<table>
<thead>
<tr>
<th>N</th>
<th>definite singular</th>
<th>definite plural</th>
<th>indefinite singular</th>
<th>bare plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>23</td>
<td>28</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>38%</td>
<td>47%</td>
<td>3%</td>
<td>13%</td>
<td></td>
</tr>
</tbody>
</table>

In (8), some examples are presented of sentences in which the verb *uitsterven* (or the adjective *uitgestorven*) is used with a bare plural in subject position.

---

5 The *INL*-corpora are made available by the Institute for Dutch Lexicology (INL). The *CONDIV*-corpus was developed within the VNC-project *Lexical variation in standard Dutch. Convergence/divergence and standardization/substandardization in Belgian Dutch and Dutch Dutch* (see Grondelaers et al. 2000).

6 Table 5.1 shows that definite plurals frequently occur in sentences in which *uitgestorven zijn* (be extinct’) or *uitsterven* (’die out’) is used. Some example sentences can be found in (i). According to Haeseryn et al. (1997: 811) this type of sentence is “dubious”. The data presented here show that there are probably speakers of Dutch who do not agree with this judgement.

Note that sentences like those in (i) indicate that definite plurals have a kind-referential interpretation (and do not unambiguously receive definite/specific readings).

(i) a. In België zijn de flankspelers uitgestorven. Spijtig, maar we in Belgium are the wingers extinct regrettable but we moeten vaststellen dat er geen meer werden opgeleid. must conclude that there no more were trained ‘In Belgium, wingers are extinct. It is regrettable, but we must conclude that no one is trained to be a winger.’

   b. Uit angst dat de Italianen uitsterven hebben de bisschoppen out of fear that the Italians die out have the bishops gepleit voor gezinnen met een derde kind. advocated for families with a third child ‘Out of fear that Italians die out, the bishops have advocated for families with a third child.’

   c. Zo dachten we dat de oesters zouden uitsterven. among other things thought we that the oysters would die out ‘Among other things, we thought that oysters would die out.’

(The sentences originate from *De Standaard*, November 1995; *NRC Handelsblad*, May 1994; and *NRC Handelsblad*, October 1994.)
(8)  
a. Bevers waren toen namelijk helemaal uitgestorven in ons land.  
Beavers were then completely extinct in our country  
‘The fact is that in those days beavers were entirely extinct in our country.’  
b. Wilde otters zijn in ons land uitgestorven door de  
Wild otters are in our country extinct through the  
 milieuvervuiling.  
‘Wild otters are extinct in our country through environmental pollution.’  
c. Mammoeten en dinosaurussen zijn uitgestorven.  
‘Mammoths and dinosaurs are extinct.’  
d. Haaien dreigen uit te sterven door overbevissing (...), zegt de  
sharks threaten out to die through overfishing (...), says the  
organisatie World Wide Fund for Nature WWF.  
organization world wide fund for nature wwf  
‘“Sharks are threatened to die out through overfishing”, says the WWF.

The set of example sentences in (8) is rather homogeneous. In every sentence, a noun phrase with an animal name is used in subject position of uitsterven/uitgestorven. So, we have to reckon with the possibility that these sentences are only representative of a particular type of language use or are only appropriate in specific contexts or registers. In (9) and (10), some additional corpus examples are given, showing that other kind predicates like voorkomen (‘be found, occur’) and (met uitroeiing) bedreigd worden/zijn (‘be threatened (with extermination)’) are also compatible with bare plurals.

(9)  
a. IJsberen komen voor op de Noordpool en in aangrenzende streken,  
Polar bears occur on the North Pole and in neighbouring areas,  
met name langs de kusten.  
especially along the coasts  
‘Polar bears are found in the Arctic and in neighbouring areas, especially along the coasts.’  
‘Bears occur in Europe, Asia, North and South America.’

---

8 The judgement that voorkomen is a kind predicate is subject to inter-speaker variation. One informant reports that according to her voorkomen does not necessarily select a kind-referring noun phrase.  
a. Tot enkele jaren geleden waren ijsberen met uitroeiing bedreigd.\textsuperscript{10} until some years ago were polar bears with extermination threatened
‘Until some years ago, polar bears were threatened with extermination.’

b. Ook door de verstoring van hun leefgebieden worden Also by the disruption of their habitats are
roofdieren bedreigd. predators threatened
‘Predators are also threatened by the disruption of their habitats.’

In (11), some examples are presented of sentences in which non-biological bare common noun phrases are used in subject position of kind predicates\textsuperscript{11}:

a. Kreidlers en Nortons waren hier bijna uitgestorven.\textsuperscript{12} kreidlers and nortons were here nearly extinct
‘Kreidlers and Nortons are nearly extinct here.’

b. Roodharige vrouwen zijn met uitsterven bedreigd. red-haired women are with extinction threatened
‘Red-haired women are threatened with extinction.’

c. Films over schaken en schakers zijn zeldzaam. ‘Movies about chess and chess players are rare.’

d. Onze korte rondgang langs de zusters (…) heeft duidelijk our short tour along the sisters (…) has clear
gemaakt dat zij het heel wat moeilijker hebben gehad dan made that they it quite a lot difficult-COMP have had than de broeders. Zusters zijn zeldzaam, dus er valt nog heel wat the brothers sisters are rare so there is still quite a lot achterstand in te halen. arrear up to make
‘Our short tour of talks with the sisters (…) has made clear that they have had a rougher time than the brothers. Sisters are rare, so there are still quite a lot of arrears to be made up.’


\textsuperscript{11} Carlson (1977), among others, treats predicates like \textit{zeldzaam} as kind predicates, although such “quantificational predicates” are not (fully) ill-formed with indefinite singular subjects (see Krifka et al. 1995:95). Importantly, most of the example sentences presented in (8), (9), (10) and (11) do not contain quantificational predicates. See chapter 8 for more discussion.

We conclude that although sentences like (8), (9), (10) and (11) are not the most frequent type of sentence, they occur in performance data. What can be derived from this evidence? The fact that such sentences are actually used by speakers of Dutch indicates that there is a number of speakers who judges such sentences to be acceptable. Corpus data do, however, not allow us to draw reliable conclusions about how many speakers judge a sentence type to be acceptable. Reliable conclusions about percentages of speakers sharing a judgement can be reached only by carrying out (questionnaire) studies into the intuitions of a number of informants. We cannot conclude from the corpus data presented in table 5.1 that the judgements in the literature are inaccurate. What we understand from the conflicting intuitions from the literature is that there is inter-speaker variation: some speakers of Dutch accept sentences like (6), whereas other speakers judge them unacceptable.

5.3.2.2 Definite plurals in characterizing sentences

This section is devoted to a second corpus study, in which we investigate the frequency of definite singulars, definite plurals, indefinite singulars and bare plurals in subject position of characterizing sentences. Recall that sentences like (7), repeated as (12), are subject to inter-speaker variation. Some speakers judge such sentences acceptable. So, we predict that this type of sentences can be found in corpora. We will verify whether this prediction is borne out.

(12) a. % De olifanten hebben waardevolle tanden.
   b. % De Amerikanen houden van glamour.
   c. % De paarden zijn zoogdieren.
   d. % De zebra’s zijn gestreept.

The results presented in table 5.2 are based on two corpora:

- A corpus of approximately 29 million words, selected from the *INL 38 Million Words Corpus* and the *CONDIV-corpus* (see for details 5.3.1.1). The texts originate from four Dutch newspapers (*Meppeler Courant, NRC Handelsblad, De Telegraaf, De Limburger*, approximately 14 million words) and four Flemish newspapers (*De Standaard, Gazet van Antwerpen, Het Laatste Nieuws, Het Belang van Limburg*, approximately 15 million words). The texts were originally published in the period 1990-2000.

- A thematic corpus of about 110 000 words. The corpus was used in Oosterhof (2003). Oosterhof (2003) presents the results of a corpus study into the relation between taxonomic hierarchies and genericity. The corpus consists of texts about polar bears, bears, carnivores and mammals. This material was selected from web sites, books,
encyclopaedias and scientific papers. The total number of words is about 110 000. The
texts were published in the period 1959-2003.

The results can be found in table 5.2.

Table 5.2: The frequency of four types of count noun phrases in subject position of
characterizing sentences

<table>
<thead>
<tr>
<th></th>
<th>definite singular</th>
<th><strong>definite plural</strong></th>
<th>indefinite singular</th>
<th>bare plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>106</td>
<td>31</td>
<td>43</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>28%</td>
<td>8%</td>
<td>12%</td>
<td>52%</td>
</tr>
</tbody>
</table>

Some examples of characterizing sentences with definite plurals in subject position are
presented in (13). These sentences express generalizations about respectively bonobos, mute
swans, polar bears and Japanese (at least in one of their readings). Notice that there is no
reason to assume that the subject noun phrases in these sentences receive taxonomic
interpretations (cf. section 3.5).

(13) a. Ellen Van Krunkelsven met een bonobo: << De bonobo’s kennen
ellen van krunkelsven with a bonobo: the bonobos know
gene machogedrag. Bij hen ligt de macht bij de wijfjes.>>
no macho attitude by them lies the power by the females
‘Ellen Van Krunkelsven with a bonobo: <<Bonobos do not have macho attitudes.
Females are in power over them.>>’

b. (…) Nu heb ik al ettelijke stervende zwanen meegemaakt,
(...) now have I already several dying swans witnesses
maar de eerste toon moet ik daarbij nog horen. Zwanen beschikken
but the first tone must i thereby still hear swans dispose
over weinig vocale mogelijkheden en vooral de knobbelzwanen
of few vocal possibilities and especially the mute swans
zijn niet bijzonder begaafd.
are not particularly talented
‘In fact I have already witnessed many dying swans, but I still have to hear the first
sound. Swans dispose of few vocal possibilities and especially mute swans are not
particularly talented.’

c. De Japanners zijn gek op walvissenspek.
the japanese are fond of blubber
‘Japanese are fond of blubber.’

---

13 The sentences originate from: Het Laatste Nieuws, 24/10/1998; Meppeler Courant, August 1993;
Greenpeace. Kwartaalblad voor donateurs. Stichting Greenpeace Nederland, 1994, vol.1; Burton, Maurice &
Amsterdam/Brussel.
d. Dit alles maakt de ijsbeer tot een onvermoeibare zwemmer.
this all makes the polar bear to a indefatigable swimmer
De lengte bereikt af en toe 2,2-80 m en het gewicht
the length reaches now and then 2-2.80 m and the weight
schiomelt, (...) tussen 400 en 700 kg. De ijsberen
fluctuates, (...) between 400 and 700 kg. the polar bears lead
leiden een uitgesproken zwervend leven en soms worden ze
an undeniably roving life and sometimes are they
ontdekt op ijsschotsen die 300 km ver in zee drijven.
discovered on ice floes that 300 km far in sea drift
‘All this makes the polar bear an indefatigable swimmer. The length sometimes
reaches 2-2.80 m and the weight fluctuates, (...) between 400 and 700 kg. Polar bears
undeniably lead a roving life and they are sometimes discovered on ice floes drifting
300 km out to sea.’

The sentences in (13) (and table 5.2) show that definite plurals can be used under
characterizing readings. This can be interpreted as evidence that there are speakers of Dutch
who find such sentences acceptable. Again, we cannot conclude from this that the well-
formedness judgements in the literature are inaccurate. The results presented in this section
only show that judgements in the literature are not necessarily representative of a language as
a whole.

5.3.3 Limitations and drawbacks of corpus based approaches

5.3.3.1 Subjective aspects of corpus research

In 5.2, we have claimed that some of the arguments that have been made against intuition-
based research apply to corpus-based approaches as well. A widespread criticism of intuition-
based research is that it is highly subjective and biased by the view of the linguist (cf. section
5.2.2.1, 5.2.3 and Schütze 1996: 1). However, much corpus-based research is biased by the
view of the linguist as well. We will argue that the decision that a corpus sentence receives a
certain interpretation is, at least to some extent, subjective (cf. 5.3.3.1.1). The same holds for
the decision that a corpus example is ill-formed and should not be included in the results (cf.
5.3.3.1.2).
5.3.3.1.1 Judgements about the interpretations of corpus sentences

In section 5.2.2.1, we cited Noël (2003: 8), who states that judgements about the interpretations of sentences (or constituents) can be very idiosyncratic. We claimed that the same problem arises if one wishes to investigate correspondences between form and meaning by carrying out corpus studies. This claim can be illustrated with an example. In Oosterhof (2005), I present the results of a corpus study into the syntactic differences between ordinary partitives and faded partitives (cf. also De Hoop, Vanden Wyngaerd & Zwart 1990). An example sentence (from De Hoop 1998) is given in (14). This sentence is ambiguous between an ordinary and a faded partitive reading. Under the latter interpretation, partitive noun phrases can be paraphrased by means of a plural noun preceded by the embedding determiner (i.e. *drie* (‘three’) in (14)) (cf. Broekhuis, Keizer & Den Dikken 2003: 556-557).

(14) Els at drie van die smerige bonbons.
    els ate three of those filthy bonbons.
    ‘Els was eating three of those filthy bonbons.’ (ordinary partitive reading)
    ‘Els was eating three filthy bonbons’. (faded partitive reading)

An important question is: how does the corpus linguist know that a corpus sentence has a certain interpretation? In many cases, the pragmatic context of an utterance is an important indication, but the syntactic properties of sentences are also relevant. The relevance of syntactic properties can be illustrated by the example sentence in (15). If a partitive construction is not introduced by an (overt) determiner, the noun phrase usually\(^\text{14}\) receives a faded partitive reading. For example, *van die mensen* in (15) unambiguously receives the faded partitive interpretation.

(15) Je vindt ze overal, van die mensen die zich op de een of andere wijze onderscheiden van anderen.\(^\text{15}\)
    ‘You can find them everywhere, people who are different from others in one way or another.’

---

\(^{14}\) There is a very restricted group of verbs which may appear as transitives or as taking a prepositional phrase with *van* (‘of’) instead of a direct object. The verbs *eten* (‘to eat’), *drinken* (‘to drink’) and *nemen* (‘to take’) are examples of this group of verbs (cf. Hoeksema 2003). This means that sentences such as (i) are ambiguous between a faded partitive and a reading corresponding to the special partitive construction described by Hoeksema (2003).

(i) Els at van die bonbons.
    els ate of those bonbons
    ‘Els ate bonbons.’ (faded partitive reading)

\(^{15}\) The example originates from *Meppeler Courant*, January 1994.
However, we have to be very careful in using this syntactic criterion when investigating the syntactic differences between ordinary partitives and faded partitives. The use of this criterion has consequences for the conclusions that can be drawn from the study. One of the results of the corpus study presented in Oosterhof (2005) is that 73% of the faded partitive constructions are not introduced by an overt determiner, while ordinary partitives always occur with an overt embedding determiner. So, one could draw the conclusion that faded partitives can, while ordinary partitives cannot occur without an overt embedding determiner. Of course, this is an example of a circular conclusion. It is more accurate to base this particular conclusion primarily on intuition-based evidence.

In a number of cases, the syntactic context does not disambiguate between the two readings. Sentences such as (14) are in principle ambiguous. In some cases where the syntactic context does not disambiguate between the two readings, the pragmatic context provides information about the interpretation of the sentence. In Oosterhof (2005), 4% of the corpus sentences is judged to be ambiguous between a faded and an ordinary partitive reading. In the relevant cases neither the syntactic nor the pragmatic context disambiguates between the two readings. The crucial point to be made is that in such studies, the corpus researcher has to decide whether the sentence is ambiguous or has one of the relevant interpretations. Such decisions are to a certain degree subjective. The reader has to trust the primary intuitions of the researcher. There is no consensus on the methodological criteria to draw conclusions about the interpretation (or ambiguity) of corpus sentences.

5.3.3.1.2 Ill-formed corpus sentences

A corpus does not provide grammaticality judgements. The fact that a sentence is taken from a corpus is not a proof of the well-formedness of the sentence (cf. McEnery & Wilson 2001 and Meurers 2005). This can be illustrated by the following example. According to the literature sentences like (16), in which a kind predicate is attributed to an indefinite singular subject, are ill-formed (cf. for example Krifka et al. 1995, Haeseryn et al. 1997 and Broekhuis, Keizer & Den Dikken 2003 and see chapter 3).

(16) # Een walvis is bijna uitgestorven
    a whale is nearly extinct

In 5.3.2.1, we have presented the results of a corpus study into the frequency of common noun phrase types in subject position of sentences in which the verb uitsterven appears. On the basis of the judgements in the literature, we can formulate the hypothesis that sentences such
as (16) do not occur in corpora. However, the fragment in (17) contains an example of a kind predicate which is predicated of an indefinite singular.\(^\text{16}\)

(17) (…) onlangs is er een toeristische boycot afgekondigd tegen Noorwegen. (…) De reden is bekend: Noorwegen (…) weigert de internationale verdragen te ondertekenen die de walvisvaart verbieden. Zo'n boycot roept een gevoel op dat verdacht veel op nostalgie lijkt. Denk aan de jaren zestig en zeventig en het lijstje van verboden vakantielanden ontrolt zich als vanzelf in je hoofd. (…) En nu dan Noorwegen. Het is geen toeval dat het om dieren gaat. Een walvis is weerloos en bijna uitgestorven en wie voor zijn lot opkomt, raakt niet in allerlei onoplosbare morele dilemma's verstrikt; die paar werkeloze vissers kunnen wel omgeschoold worden.\(^\text{17}\)

‘Recently, a tourist boycott of Norway was announced. (…) The reason is known: Norway (…) refuses to sign the international treaties that prohibit whale fishing. Such a boycott evokes a feeling that suspiciously resembles nostalgia. Think about the sixties and seventies and a list of forbidden countries unfolds in our minds. (…) And now Norway. It is no coincidence that animals are involved. Whales are helpless and nearly extinct and who stands up for their fate does not get caught up in all kinds of unsolvable moral dilemmas; the small number of unemployed fishermen can be retrained.’

Importantly, (many) native speakers of Dutch deem the underlined sentence in (17) ill-formed. The sentence clearly conflicts with our linguistic competence. This illustrates that performance data give a distorted picture of I-language. Other examples of ill-formed corpus sentences can be found in Van der Beek (2005: 3) and Colleman (2006: 226-227).

A possible methodological approach is to filter out ungrammatical sentences (or to extract only the grammatical ones). This procedure is used by Van der Beek (2005) (cf. also Colleman 2006 for some discussion). Van der Beek (2005: 3) writes:

“In the evaluation step, the good examples are extracted from the total set of candidates. (...) This set will (...) contain sentences which (...) do not contain the linguistic structure under investigation. (...) On top of that, the corpus is likely to contain (...) ungrammatical sentences.”

In table 5.1 we indicated that in one corpus sentence an indefinite singular occurs in subject position of uitsterven. The relevant sentence is the one in (17). If we had based our results on the set of “grammatical sentences”, the table would have looked as follows:

\(^{16}\) It is, of course, not a coincidence that the indefinite singular occurs in subject position of a coordinate structure, in which the first conjunct corresponds to a characterizing interpretation. Possibly, sentences such as (16), which does not contain coordination, do not occur in corpora. This does, however, not contradict the fact that the judgements in the literature predict the corpus sentence in (17) to be ill-formed.

\(^{17}\) The sentence originates from *NRC Handelsblad*, March 1994.
I will discuss some drawbacks of this method. Under this methodology, the results of the corpus analysis neither reflect linguistic competence nor performance. The results are based on a collection of performance data from which sentences that are not in accordance with the linguistic competence of the author (and possibly other language users) are removed. Conclusions about (among other things) the well-formedness of sentence types are obscured if “ungrammatical” sentences are filtered out in an earlier stage of the research process. Newmeyer (2003: 702) puts it as follows: “grammar is grammar and usage is usage.” Corpus research is a method for investigating usage. Data about usage can be interpreted as indirect evidence about grammar, but data about usage may not be ‘mixed’ with data about grammar.

There could be some legitimate methods to objectify the methodological decision that a sentence is ill-formed. One possibility is that the researcher includes an at first sight “wrong” construction only if it appears several times. Another is that we use dictionaries and grammar books to verify whether a construction is mentioned. However, such a methodology raises questions like: How many attestations in how many different resources are enough to conclude that something is not an accidental slip of the pen? In how many dictionaries or other resources must a (dubious) construction be mentioned in order to be included in the results of the study? It is in principle impossible to answer these questions satisfactorily, because there is no one-to-one relation between, on the one hand, well-formedness and, on the other hand, the frequency of a construction in a corpus or the fact that it is mentioned in a dictionary or a grammar.

Because of these disadvantages and problems, ill-formed sentences have not been filtered out from the set of sentences on which the results of the corpus studies presented in this thesis are based. The corpus results presented in this study are purely based on performance data.

5.3.3.2 The absence of sentence types

A second drawback of corpus research has to do with the fact that the set of well-formed sentences of a language is in principle infinite (cf. for example Chomsky 1957). However, the size of a corpus is restricted. This raises the following question: “How can a partial corpus be the sole explicandum of an infinite language?” (McEnery & Wilson 2001:12). In broad terms, the result of a corpus study is always a limited set of relevant sentences. We do not know whether a sentence type that was not found is unacceptable or did not occur in the corpus by
coincidence. Again, we conclude that there is still an important role for acceptability judgements (cf. also Borsley & Ingham 2002: 5).

This fact has implications not only for syntactic studies but also for semantically oriented research. Suppose that a corpus does not contain sentences in which an indefinite singular is used in subject position of a kind predicate (cf. (16)). From this result alone we cannot conclude that such sentences are unacceptable. Whether a sentence is unacceptable is determined by the judgements of language users.

The same point can be made by referring to the results of a corpus study into the distribution of a negative polarity item (cf. Hoeksema 2004, cited in Oosterhof 2003-2004). The expression *het feest gaat door* (‘the event will take place’), in which *feest* (‘party’) is to be interpreted metaphorically, is an example of an idiom that is sensitive to polarity. Hoeksema’s material (cf. Hoeksema 2004) contains 56 examples of this idiom. In all these sentences, the trigger *niet* (‘not’) occurs (cf. (18) for a sentence with *niet*).

(18) Liefst 115 miljoen dollar wilde Ellison schenken aan Harvard University, as much as 115 million dollar wanted ellison donate to harvard university maar helaas, *het feest gaat niet door*.18 but unfortunately the party goes not through ‘Ellison wanted to donate as much as 115 million dollars to Harvard University, but this plan does not go through.’

The fact that in each of Hoeksema’s corpus sentences the trigger *niet* (‘not’) occurs is an important indication that *het feest gaat door* is a negative polarity item. It is, however, known from the literature that many negative polarity items do not occur exclusively in purely negative contexts. Giannakidou (1998) argues that an item is a polarity item if it occurs in (a subset of the set of) “nonveridical” contexts (cf. also Zwarts 1995). A context is “nonveridical” if and only if the event described in the context is not asserted to occur (see Giannakidou 1998 for further details). This notion can be illustrated by sentence (19a). The conditional sentence *als dat feest doorgaat* is a nonveridical context, since sentence (19a) does not entail that the event (or situation) referred to by *dat feest* takes place. The same holds for sentence (19b), in which the intensional verb *hopen* (‘to hope’) is used. These sentences are well-formed.

18 This sentence originates from http://www.pmmblognoot.nl/2006/06/larry_ellison_t.html (August 2006).
(19)  a. Als dat feest doorgaat, slaan we een flinke slag
    ‘If that event will take place, we make a good deal.’

    b. Ik hoop dat het feest doorgaat.
    ‘I hope that the event will take place.’

Van der Wal (1996) observes that certain negative polarity items are now and then encountered in sentences that are emphatically affirmative, due to the presence of (stressed) adverbs such as wel (‘indeed’). An example sentence is presented in (20). (This illustrates that Giannakidou’s view may be too narrow.)

(20)  Dat feest gaat WEL door
    ‘That event will take place indeed.’

Recall that sentences such as (19) and (20) do not occur in Hoeksema’s corpus. However, we cannot conclude from this fact that such sentences are ill-formed. The absence of such sentences in a corpus does not indicate that they are unacceptable. If we want to verify whether het feest gaat door can occur in contexts like the ones illustrated in (19) and (20), we have to make use of introspection-based data.

Our conclusion must be that corpus studies do not provide a complete picture of semantic (and other) phenomena. Linguistic intuitions continue to play an important role in interpreting corpus results and describing the semantics of natural language.

5.3.4 Corpus-based approaches in this thesis

In the previous sections, we have discussed that it is difficult to draw reliable conclusions about acceptability on the basis of corpus-based approaches. We have observed that the fact that a sentence occurs in a corpus does not guarantee that the sentence is acceptable and that the fact that a sentence does not appear in a corpus does not guarantee that it is unacceptable. This is not to say that there is no relation at all between corpus results and acceptability judgements. A plausible hypothesis is that if a sentence type occurs frequently in a corpus, there are a number of speakers who judge the sentence type acceptable. It is, however, not clear how frequently it has to occur before we can draw this conclusion. Another plausible hypothesis is that if a sentence X is judged acceptable by more speakers than a sentence Y, there is a good chance that sentence X occurs more frequently in corpora. Corpus data about a
sentence type do, however, not justify conclusions about the (exact) number of speakers judging the sentence acceptable.

Because of the limitations of corpus research, (many of) the conclusions we will draw about acceptability and variation among speakers are verified by intuition-based approaches. The importance of corpus-based approaches to this thesis lies in the fact that corpus studies can reveal frequency differences between sentence types, which can be relevant to theoretical proposals.

5.4 Summary and conclusions

In this chapter, we have discussed the empirical basis of semantic research. We have reviewed the merits and problems of introspection-based and of corpus-based approaches.

Judgements about the interpretation and the acceptability of sentences are, at least to a considerable extent, subjective. The intuitions of one speaker are not necessarily representative of the language. They are representative of the I-language of an individual speaker. We have argued that the subjectivity of intuitions is a welcome aspect of intuition-based research: the subjective intuitions of native speakers of a language are an appropriate method for investigating inter-speaker variation.

Corpus results often do not tell us anything about inter-speaker variation. They do not provide an appropriate basis for drawing conclusions about how many (and what) speakers share a judgement. Another limitation is that there is no one-to-one relation between the occurrence of a sentence type in a corpus and the acceptability of the sentence type. The fact that a sentence occurs in a corpus does not guarantee that the sentence is acceptable and the fact that a sentence does not appear in a corpus does not guarantee that it is unacceptable. We have noted that there are some situations in which we can draw conclusions about acceptability. For example, if a sentence type occurs frequently in a corpus, we have an indication that there are speakers who judge the sentence type acceptable. But in general, the relevance of corpus approaches to the issue of acceptability and variation among speakers is limited. In the present study, (many of) the conclusions about acceptability and variation are verified by investigations of the judgements of groups of speakers.