Why I will not become a corpus linguist

The use of introspection data and corpus data in synchronic syntactic research

Hans Broekhuis (Meertens Institute, Amsterdam)

1 Introduction

Not so long ago it seemed unnecessary to motivate the use of what is nowadays called the intuitionist approach to the study of syntax: the study of the internal structure of phrases/sentences with the help of data obtained by introspection. But times have changed and the reviews of the *Syntax of Dutch* (2012-2019) that appeared in earlier issues of *Nederlandse Taalkunde* leave no doubt that some of the reviewers consider this approach to linguistics inadequate (if not obsolete): a reference grammar such as *Syntax of Dutch* should be based on corpus data. The presupposition underlying this claim seems to be that introspection data are not empirical data and are therefore inherently inferior to corpus data based on "real language" (in the words of De Hoop 2016). In my rejoinder to these reviews (Broekhuis 2016), I think I made it clear that the reviewers in question were overly optimistic about the potentials of corpus linguistics. At least some of the improvements and corrections they proposed did not hold water, due to the fact that they were based on using "raw" corpus data or that there were shortcomings in the linguistic annotation of the corpora used.

Fortunately, not all linguists who promote the use of corpora are of the same opinion as the reviewers in question: Odijk (to appear), for instance, stresses that data from corpora and data collected in artificial experimental settings (*including* introspection) should all be considered empirical data: "all relevant evidence should be taken into account and no form of evidence has a privileged status." Although I agree with Odijk's statement in principle, this article will argue that there are reasons for assuming that introspection research is a better method for collecting synchronic syntactic data than corpus research. Note that other linguistic domains (like diachronic syntax, language variation and language acquisition), where corpus research is often indispensable in finding relevant data, are left out of account here.

Not much of what follows is new; see Newmeyer (1983) and Kübler and Zinsmeister (2015:17) for in-depth discussions of data collection by means of introspection and corpora from two different angles. I believe, however, that this article is justified here as at least some linguists objecting to the intuitionist approach seem insufficiently informed about the nature of synchronic syntactic research and the vices and virtues of the two methods of collecting data in relation to it. Furthermore, the arguments against the intuitionist approach to synchronic syntax I have seen so far strike me as being merely dismissive in nature, while I hope that this article will help critics of this approach in considering more informed arguments against it, which may be more conducive to scientific debate.

2 Competence and performance

This article investigates the nature of the data needed for studying synchronic syntax (the internal structure of phrases/sentences of present-day languages) and what it tells us about the adequacy of the two methods of collecting data under discussion. Studying the internal organization of phrases/sentences is not an easy task and progress has been slow but steady since the introduction of generative grammar in the mid 1950's. The fact that the complexity of syntactic structures is not immediately evident from actual utterances but can only be brought to light in an indirect way by applying a series of complicated and sometimes controversial syntactic tests is taken in generative grammar as evidence for postulating a certain amount of tacit knowledge in the mind/brain of the language user enabling him to produce and process these structures. Although relatively little is known about the form of syntactic knowledge in the mind/brain, a great deal is known about the presumably universal principles determining what are possible/grammatical and impossible/ungrammatical syntactic structures. The set of universal principles determining syntactic structure is taken to be the innate part of the "internal language" or the "competence" of the adult language user that determines the properties of his language, and is also taken to be instrumental in language acquisition. If true, studying the syntax of a specific language L is not only of interest in its own right but also for the study of the universal mechanisms underlying the ability of humans to produce, process and acquire language.

Some linguists deny the existence of competence in the Chomskian sense: a relatively stable knowledge state that arises in the language-learning child at a fairly early age (say, before adolescence). Usage-based theory, for instance:

> […] incorporates the basic insight that usage has an effect on linguistic structure. It thus contrasts with the generative paradigm's focus on competence to the exclusion of performance and rather looks to evidence from usage for the understanding of the cognitive organization of the grammar" (Bybee & Beckner, 2010).

It should be noted, however, that even if the "basic insight that usage has an effect on linguistic structure" were beyond doubt, this need not imply that the universal principles determining syntactic structure detected by generative research should be abandoned, as these are claimed to hold for any syntactic knowledge system of adults about their native language, regardless the question of whether it should be considered a stable or a more flexible knowledge state: if it were flexible, the universal principles would simply hold for all consecutive stages. In view of the empirical success attributed to universal principles by generative linguists, dismissing these principles by simply pointing to the presumed flexible nature of the grammatical system alone, as done in Van de Velde (2014:89), is clearly insufficient; one would at least expect an attempt of refuting these principles on empirical grounds. While there are many generative studies attempting this as part of the normal scientific practice of theory evaluation, I do not know of any user-based study that specifically aims at this. I wish to maintain that postulating such universal principles as an innate part of the knowledge state of language users is empirically grounded, which entails that the distinction between competence and performance is also meaningful for individual speakers: *competence* refers to the mechanisms within the individual enabling him to produce and process the abstract linguistic objects that generative grammar refers to as phrases/sentences; performance refers to everything related to the use of utterances built on these abstract objects. Because it is not important for the present discussion what the substantial content of competence is, the reader may just think of it in a fairly theory-neutral way as a set of hypotheses about the internal structure of phrases/sentences.

3 What kind of data is needed for competence research?

Competence enables the individual speaker to produce and process structured linguistic objects. Syntactic research aims at modeling the adult speaker's knowledge stage, and by comparing the knowledge states of larger groups of speakers, at uncovering the universal principles that enable the child to acquire natural language. Reaching these goals is a step-by-step process: on the basis of the analysis of a necessarily limited number of syntactic phenomena in a limited number of languages, hypotheses are proposed concerning the universal principles that linguistic objects subject to. These universal principles may affect the analysis of the original data set by making a selection from the competing analyses that can in principle account for this set. They may also make predictions about other, unrelated syntactic phenomena and the syntactic behavior of languages not considered before. This leads to a gradual extension of the empirical coverage, which may give rise to revised or new hypotheses about the universal principles. By going through this process over and over again, we hope to eventually increase our insight in the universal properties of the competence of speakers of different languages.

The data taken into account for competence research can be selected rather randomly: any set that the researcher considers eligible for fruitful investigation will do. There are however, two important notions that restrict the data set, namely acceptability and grammaticality; cf. Newmeyer (1983: §2.2). The notion *grammaticality* is the easier one to define, as it is a theoretical term: a syntactic object is grammatical if the language user is predicted to be able to produce it, and ungrammatical if the language user is predicted not to be able to produce it. The term is still problematic, however, because competence theory is continuously being updated, which entails that certain objects can be characterized as grammatical at one stage of the theory and as ungrammatical at another stage (and vice versa). The notion of *acceptability* is not a theoretical term but refers to the feelings/intuitions that language users have about utterances. These intuitions depend on the speaker's competence, but are also affected by numerous other factors like interpretability and language norms. Acceptability has little to do with actual use, as speakers may in fact use utterances that they judge unacceptable.

At first sight, the two notions may not seem very helpful in restricting the data set, but there seems to be a consensus that at least those utterances that are produced frequently in colloquial speech and are judged acceptable by native speakers should be contained in the set of grammatical sentences, and that utterances that occur infrequently (or never at all) in colloquial speech and are judged unacceptable by native speakers should be contained in the set of ungrammatical sentences. We can illustrate this as in Figure 1 by considering two factors that affect acceptability of linguistic forms: grammaticality and interpretability. These factors divide any data set in four, possibly empty, subsets (if we ignore, for simplicity's sake, the many other factors that may affect acceptability judgments):

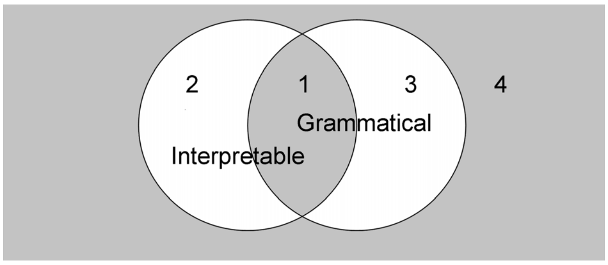

Figure 1: Two factors determining acceptability

The nature of subset 1 (grammatical and interpretable) will probably be unproblematic for most readers, as these are the forms that all native speakers will characterize as acceptable, that is, as belonging to their language; we may therefore expect these forms to occur in speech produced by native speakers. The same will hold for subset 4 (ungrammatical and uninterpretable), as these are forms that native speakers will all characterize as unacceptable; we may expect these form not to occur in the speech of native speakers (although they may occur in the speech of, e.g., non-native speakers). Subset 2 (interpretable but ungrammatical) is a mixed bag, which may contain evident speech errors such as *De boek is mooi* ‘The book is beautiful’, where the neuter noun *boek* is preceded by a non-neuter article *de*. Subset 3 (grammatical but uninterpretable) is more problematic as the decision to locate a certain form in this subset depends on the grammar that is assumed: whether Chomsky’s famous example *Colorless green ideas sleep furiously* belongs to this set, for instance, depends on the question whether the grammar includes selection restrictions such “the verb *sleep* takes a human subject”: if it does, this sentence belongs to subset 4, but if it does not, the sentence belongs to subset 3. There are many forms, for which it may be difficult to decide where they belong. For instance, native speakers of Dutch occasionally place pseudo-participles (adjectives with the appearance of a participle) such as *geliefd* ‘popular’ in the position following verbs in clause final position, as in (1b).

(1) a. Marie zegt dat zijn boeken geliefd zijn.
Marie says that his books popular are
‘Marie says that his books are popular.’
b. *Marie zegt dat zijn boeken zijn geliefd.

My own opinion (expressed by the use of an asterisk) is that (1b) belongs to subset 2, while others may maintain that it belongs to subset 1, or illustrates some ongoing language change. Since it is virtually impossible to provide a conclusive argument for any of these positions, this has led to the methodological guideline that such examples should not be used for evaluating theories, but to let the prevailing theory of the time decide whether the example should be considered grammatical or not. The theory will thus be based on clear (undisputedly grammatical/ungrammatical) cases only, which avoids *ad hoc* additions to the theory or elimination of otherwise well-motivated mechanisms from the theory on shaky grounds. This means that competence research mainly avails itself of the clear cases from subset 1 and 4 (the grey area in Figure 1), and excludes examples from subset 2 and 3 from the data set used for theory evaluation. Until now, the set of clear cases (for a large part still to be explored) has provided sufficient material for new and challenging research and this will be the case for a long time to come. We may hope that, by the time that the set of clear cases is exhausted, we may be informed enough to also say something sensible about the less clear cases although I am not very optimistic in this respect.

4 How the intuitionist researcher compiles the data set

This section illustrates that compiling a data set by means of the intuitionist approach is normally a fairly organized process by using (part of) the data set collected in Broekhuis and Den Dikken (2018), henceforth B&DD. Their investigation starts with the simple observation that the PP *Tot aan$_1$ het einde aan$_2$ toe* 'up to the end' can be reduced by omission of *aan$_1$*, *aan$_2$* and/or *toe*. By systematically going through all the options, we are able to construct the data set in (2). The primeless examples to the left show that *aan$_2$* and *toe* are both optional but that *toe* must be present when *aan$_2$* is overtly realized; the contrast between the primeless and primed examples show that *aan*$_1$ is optional in all cases. Exploiting the standard use of parentheses (indicating optionality), we can collectively refer to all and only the acceptable forms in (2) by means of the string *tot (aan) het einde ((aan) toe)*, which was used as the main title of B&DD.

(2) a. tot het einde
to the end

b. tot het einde toe
to the end to

c. tot het einde aan$_2$ toe
to the end on to

d. *tot het einde aan$_2$
to the end on

a′. tot aan$_1$ het einde
to on the end

b′. tot aan$_1$ het einde toe
to on the end to

c′. tot aan$_1$ het einde aan$_2$ toe
to on the end on to

d′. *tot aan$_1$ het einde aan$_2$
to on the end on

B&DD noted that there are subtle meaning differences between the acceptable forms in (2), taking this as evidence for the claim that the functional make-up of these forms differ. I will not repeat the whole article here, but confine myself to giving a number of empirical facts that are crucial for their analysis. B&DD started by showing that the grammatical forms in (2) can be used as adverbial phrases of time and place only, as illustrated in (3a) and (4a). The first crucial observation is that the preposition *tot* in these examples cannot be omitted; when *toe* is present, the result is always unacceptable, as indicated by the asterisk in the (b)-examples; when *toe* is absent, the result is unacceptable or results in the loss of the intended 'up to'-interpretation, as indicated by the hash sign in the (c)-examples.

(3) a. Jan heeft tot (aan) het einde ((aan) toe) geslapen. [time]
Jan has to on the end on to slept
'Jan has slept up to the end (of e.g. the meeting).'
b. *Jan heeft (aan) het einde (aan) toe geslapen.
c. Jan heeft *($^{\#}$aan) het einde geslapen.

(4) a. Jan heeft het gras tot (aan) het einde ((aan) toe) verwijderd. [spatial]
Jan has the grass to on the end on to removed
'Jan has removed the grass up to the end (e.g. from the garden path)'
b. *Jan heeft het gras (aan) het einde (aan) toe verwijderd.
c. Jan heeft het gras *($^{\#}$aan) het einde verwijderd.

Given the independently established fact that lexical heads must be overtly realized in non-ellipsis contexts (due to the principle of recoverability), the examples in (3)-(4) can be taken as evidence for the claim that *tot* is the head in all PP-constructions, which takes the remainder of the PPs as its complement, as in (5).

(5) [$_{PP}$ *tot* [*(aan) het einde ((aan) toe)*]].

Another crucial observation for their analysis is that the phrase *aan het einde* in (2a′-c′) can be replaced by the adverbial proforms *daar* 'there' and *dan* 'then'.

(6) a. tot aan het einde a′. tot daar/dan
b. tot aan het einde toe b′. tot daar/dan toe
c. tot aan het einde aan toe c′. tot daar/dan aan toe

Given the independently established fact that such replacement is possible with phrases only, the examples in (6) shows that *aan het einde* is a PP embedded in the complement of *tot*. On the assumption (motivated in the original paper) that *toe* is a postposition, this leads to the structures in the primeless examples in (7), where the PPs are numbered according to their depth of embedding. The structures in the primed examples follow on the assumption that the NP *het einde* is located in a similar position as the PP *aan het einde*.

(7) a. [$_{PP1}$ tot [$_{PP2}$ aan het einde]] a′. [$_{PP}$ tot [$_{NP}$ het einde]]
b. [$_{PP1}$ tot [$_{PP2}$ [$_{PP3}$ aan het einde] toe]] b′. [$_{PP1}$ tot [$_{PP2}$ [$_{NP}$ het einde] toe]]
c. [$_{PP3}$ tot [$_{PP2}$ [$_{PP3}$ aan het einde] aan toe]] c′. [$_{PP3}$ tot [$_{PP2}$ [$_{NP3}$ het einde] aan toe]]

In order to arrive at a full analysis of the paradigm, we still need to account for the occurrence of *aan*$_2$ in the (c)-examples and the unacceptability of the (d)-examples in (2).

The reader is referred to the article for the full analysis, as the above is sufficient for our limited goal of showing how the intuitionist researcher compiles the data set: the most important finding is, to my mind, that the research questions themselves dictate how the data set should be expanded in order to find proper answers.

5 The validity of introspection data

Section 1 mentioned that some of the critics of *Syntax of Dutch* disapprove of the way the intuitionist researcher compiles his data set: this set normally consists of introspection data, which are thought to be subjective and unreliable in that acceptability judgments can be theoretically biased. Introspection data are therefore inherently inferior to corpus data based on "real language". It is not so clear that introspection data are really as subjective and unreliable as claimed by these critics: the few available studies suggest that the reliability of data sets based on introspection is not significantly smaller than of those based on corpora: in so far as the two can be compared, convergence seems rather to be the rule than the exception (see Sprouse and Almeida 2010). The reason for this may be that most introspection data found in the literature are not truly subjective but intersubjective: when a researcher is not entirely sure about his judgments, he will normally consult others about their judgments in order to sharpen his own. Furthermore, the data are normally scrutinized during peer review before publication and may also be subject to critical investigation after publication in subsequent work on the same topic. In case doubt remains, there is always the option to investigate the debated forms by more controlled research (such as large scale intuition research). Faulty introspection data simply do not have much chance to survive in the long run and are normally already weeded out before publication of the research.

6 What kind of data are found in corpora

Since the dismissal of introspection data is not based on actual research, the suggestion that linguistic research should be based on "real language" seems to reflect a prejudice in favor of "real language". An important question is whether this prejudice can be justified by the data found in corpora, as far as synchronic syntactic research is concerned, The answer is clearly negative, as will become clear from the following three citations from Kübler and Zinsmeister's (2015:17) handbook on corpus linguistics concerning "three shortcomings of corpus data that users need to be aware of". The first citation concerns the competence - performance dichotomy:

> [...] corpus data are always *performance data* in the sense that they are potentially distorted by external factors such as incomplete sentences, repetitions, and corrections in conceptionally spoken language, or text-genre related properties in conceptionally written language such as nominalizations in law-related texts.

The competence researcher who bases himself on corpus data is thus forced to weed out the unclear data in order to meet the methodological demand that competence theory is based on clear cases (cf. §3), which can only be done be appealing to ... intuition. The second citation concerns completeness of the data set:

> [...] corpora are always *limited*; even web-based mega corpora comprise only a limited number of sentences. This means that there is a certain chance that the web pages with the linguistically interesting variant might just have been omitted by chance when the corpus was created.

The situation is actually more serious than this, as is clear from Kübler and Zinsmeister's (2015:166) addition that linguistically annotated corpora cannot be used for the search for rare examples. The competence researcher who bases himself on corpus data is thus forced to supplement the data set with the missing cases, which can only be done be appealing to ... intuition. The third citation concerns negative evidence:

> [...] corpus data will never provide *negative evidence* in contrast to acceptability judgements—even if there are ungrammatical examples in corpora. This means that corpus data do not allow us to test the limits of what belongs to the language and what does not.

Given that 'the limits of what belongs to the language and what does not' is the core business of the competence researcher, this citation indicates that the competence researcher who bases himself on corpus data can only do his job when he supplements the data set with the negative evidence, which can only be done be appealing to ... intuition.

7 The relation between competence and corpus research

The conclusion drawn from the citations from Kübler and Zinsmeister (2015:17) can only be that without an appeal to introspection data, there will be no competence research. Corpus data are unsatisfactory for this type of research by their very nature and can at best be used to adjust intuitionist claims about clear data by showing that some presumed unacceptable example occurs with a high frequency or that some presumed acceptable example does not occur at all, but even then introspection is called for. On the other hand, it is clear that intuitionist research has had a major impact on corpus research in the guise of the tag sets employed: “the most common type of linguistic annotation is part-of-speech annotation which can be applied automatically with high accuracy, and which is already very helpful for linguistic analysis despite its shallow character” (Kübler and Zinsmeister (2015:17). The part-of-speech annotation at least is based on the results of centuries of intuitionist linguist research and thus carries over all the presumed subjectivity and reliability of the intuitive approach to corpus research. The shallow character of the (partly outdated) annotation brings with it that it requires a lot of creative thinking on the part of the corpus researcher to select the data relevant for specific research questions and even then the results are questionable when it comes to the evaluation of competence research; see Broekhuis’ (2016) remarks on the corpus study in Van Bergen and De Swart (2010). These problems can only be remedied by developing more sophisticated annotation systems, which, I believe, can only be provided by ... intuitionist research.

Acknowledgment: I like to thank Frits Beukema for copy-editing this article.

About the author

Meertens institute Amsterdam
hans.broekhuis@meertens.knaw.nl
https://www.linkedin.com/in/hbroekhuis/

References

Broekhuis, Hans (2016). Syntax of Dutch: the data set. *Nederlandse Taalkunde* 21, 297-325.

Broekhuis, Hans, Norbert Corver, Marcel Den Dikken, Evelien Keizer, and Riet Vos (2012-2019). *Syntax of Dutch (8 volumes)*. Amsterdam: Amsterdam University Press.

Broekhuis, Hans, and Marcel Den Dikken (2018). *Tot (aan) het einde ((aan) toe*: The internal syntax of a Dutch complex PP. *Glossa* 104, 1-19.

Bybee, Joan L., and Clay Beckner (2010). Usage-based theory. In: Bernd Heine and Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*. Oxford, Oxford University Press, 828-855.

De Hoop, Helen (2016). Woordvolgordevariatie: theorie versus empirie? *Nederlandse Taalkunde* 21, 265-284.

Kübler, Sandra, and Heike Zinsmeister (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. London/new York: Bloomsbury Academic.

Newmeyer, Frederick J. (1983). *Grammatical theory. Its limits and possibilities*. Chicago/London: University of Chicago Press.

Odijk (to appear). De verleidingen en gevaren van GrETEL. *Nederlandse Taalkunde* 25.

Sprouse, Jon, and Diogo Almeida (2010). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48, 609-652.

Van Bergen, Geertje, and Peter de Swart (2010). Scrambling in spoken Dutch: definiteness versus weight as determinants of word order variation. *Corpus Linguistics and Linguistic Theory* 6, 267-295.

Van de Velde, Freek (2014). Nederlandse predeteminatoren als levend fossiel. *Nederlandse Taalkunde* 19, 87-103.