CLARIN-supported Research on Modification Potential in Dutch First Language Acquisition

Jan Odijk
Utrecht University, The Netherlands
j.odijk@uu.nl

Abstract

This paper analyses data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning 'very', which show syntactic differences in modification potential. It continues the research reported on in (Odijk, 2016). The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular the treebank search applications *PaQu* (Parse and Query) and *GrETEL* Version 4.00. The analysis benefits from the use of parsed corpora (treebanks) in combination with the search and analysis options offered by PaQu and GrETEL. Earlier work showed that despite little data for *zeer* modifying adpositional phrases adult speakers end up with a generalised modification potential for this word. In this paper, I extend the dataset considered, and find more (but still little) data for this phenomenon. However, I also find a similar amount of data that form counterexamples to the non-generalisation of the modification potential of *heel*. I argue that the examples with *heel* concern constructions with idiosyncratic semantics and therefore are not counted as evidence for the general rule of modification. I suggest a simple statistical analysis to account for the fact that children 'learn' that *heel* cannot modify verbs or adpositions though there is no explicit evidence for this and they are not explicitly taught so.

1 Introduction

In this paper I analyse data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning 'very'. It continues the research reported on in (Odijk, 2016). The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular the treebank search applications *PaQu* (Parse and Query (Odijk et al., 2017)) and *GrETEL* Version 4.00 (Odijk et al., 2018), both of which make use of the Dutch syntactic parser Alpino (Bouma et al., 2001). The words that are being investigated are highly ambiguous. Most of the ambiguity is resolved by considering the syntactic context they occur in. Therefore, the analysis benefits from the use of parsed corpora (treebanks). Though the automatically created parses contain errors and require manual verification, the data analysis process is considerably speeded up and facilitated by these parses in combination with the search and analysis options offered by PaQu and GrETEL.

This paper is organised as follows: I introduce the linguistic problem in section 2. Section 3 introduces the treebank search applications used. Section 4 describes earlier work done on this type of problem and on the specific problem itself. This earlier work was carried out on relatively small corpora. Section 5 describes the complexity of first language acquisition and the simplifications and idealisations I assume to address the problem. In section 6 I describe which corpora I used in the research and report on the treebank query results found. Section 7 proposes considerations that may lead to an analysis of the problem. Section 8 summarises the main findings of this paper and suggests future research.

Jan Odijk 2020. CLARIN-supported Research on Modification Potential in Dutch First Language Acquisition. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 94–107.

2 The Problem

The three Dutch words *heel*, *erg* and *zeer* are (near-)synonyms meaning 'very', i.e. (stated informally) they modify a word or phrase that expresses a (gradable) property or state and specify that its modifiee has the property or state it expresses to a high degree. Of these, *heel* can modify adjectival (A) phrases only, while *erg* and *zeer* can modify not only adjectival, but also verbal (V) and adpositional (P) phrases. This is illustrated in example (1).[1]

(1) a. Hij is daar heel / erg / zeer blij over
he is there very / very / very glad about
'He is very happy about that'

b. Hij is daar *heel / erg / zeer in zijn sas mee
he is there very / very / very in his lock with
'He is very happy about that'

c. Dat verbaast mij *heel / erg / zeer
That surprises me very / very / very
'That surprises me very much'

In (1a) the adjectival phrase *blij* 'glad' can be modified by each of the three words. In (1b) the (idiomatic) adpositional phrase (PP) *in zijn sas* can be modified by *zeer* and *erg* but not by *heel*. The same holds in (1c) for the verbal phrase *verbaast*.[2] In English, the same holds for the word *very*: it can only modify adjectives.[3] For verbs and prepositional phrases one cannot use *very* but one can use the expression *very much* instead:

(2) a. He is very happy about it
b. He is *very / very much in love with her
c. It surprised me *very / very much

The distinctions illustrated in the preceding section are purely syntactic in nature. The words *heel*, *zeer* and *erg* are synonyms or near-synonyms, and the expressions *blij* and *in zijn sas* are near-synonyms as well, which makes it unlikely that the differences can be derived from semantic properties. It is also not in any way obvious how the differences could follow from universal principles of language or language acquisition.

There are other differences among the words *heel*, *erg* and *zeer*. If any of these differences is somehow related to the difference under investigation then it must be a difference in which *heel* opposes the other two words *erg* and *zeer*. However, this is not the case (Odijk, 2015).

3 The Treebank Search Applications PaQu and GrETEL 4.0

It is important to investigate the use of these words in their syntactic context, because they are (as many words in natural language) highly ambiguous. Odijk (2016) shows that *heel* is 6-fold ambiguous, *erg* is 4-fold ambiguous, and *zeer* is 3-fold ambiguous, but he also shows that the ambiguity is largely resolved by considering the syntactic context. For this reason I address the problem using the treebank search applications *Parse and Query* (*PaQu*) (Odijk et al., 2017) and *GrETEL* Version 4.00 (Odijk et al., 2018).

Both applications make existing manually verified treebanks for Dutch such as *LASSY-Small* for written Dutch (van Noord et al., 2013) and the *Spoken Dutch Corpus* (Oostdijk et al., 2002) available for search. They also enable a researcher to upload a text corpus and associated metadata, and have it automatically parsed by the Alpino parser (Bouma et al., 2001), after which the resulting treebank is made available for search.

[1] An asterisk is used to mark ill-formed expressions.

[2] or maybe the whole VP *verbaast mij*.

[3] and certain adverbs. I assume that words traditionally assigned the part of speech 'adverb' are either adjectives or (intransitive) adpositions.

The syntactic structures inside the treebanks are encoded in XML. Both applications offer XPath to search in these syntactic structures for words, grammatical properties and constructions. Each of them also offers additional search options: PaQu offers a very easy way to query for grammatical dependency relations between words, and GrETEL offers query by example facilities (Augustinus et al., 2012). Both treebank applications also offer various ways of analysing the search results, for data and metadata combined.

4 Earlier work

The type of problem dealt with here has, at least for English phenomena, figured prominently in the language acquisition literature (Baker, 1979; Berwick, 1985; Pinker, 1989; Yang, 2016), e.g. for accounting for the acquisition of adjectives that can be used predicatively but not attributively, and for accounting for dative constructions, in which some but not all verbs allow the double object construction in addition to the *to*-dative construction. This paper will not propose a general new solution to this problem, but has the more modest aim of analysing the relevant Dutch data for the problem at hand.

Odijk (2015) analyses the Dutch CHILDES corpora (MacWhinney, 2000) for the words *heel*, *erg* and *zeer*. These corpora contain transcriptions of adult-child interaction with (monthly) sessions recorded between the children's ages of approximately 1 year and 8 months and 6 years.[4] Since the children have to acquire the lexical properties of these words from the input provided by the adults (and other participants), this work focuses on the child-directed speech. The findings, together with findings in additional corpora, will be summarized in section 6.

5 First Language Acquisition

First language acquisition is extremely complex: the input is speech, which has to be turned into a sequence of phonetic symbols by the child while it has to build up the phone(me) inventory of the language it is acquiring at the same time. The speech is spontaneous, and therefore contains phenomena that are typical for spontaneous speech such as:[5]

Hesitations and filled pauses e.g. *en ehm (.) gaan we nog ehm (.)+* (*and hmm go we still hmm*)

Repetitions *een molen [/] molen* (*a mill mill*)

False starts and retracing <geef jij> [//] kom jij op mijn verjaardag ? (*give you come you on my birthday ?*)

Unfinished utterances (see example under hesitations)

Of course, the speech signal does not contain word boundaries, so the child has to find out somehow where the word boundaries are so that the input sequence of phone(me)s can be tokenized into a sequence of word tokens.

For the phenomenon under investigation here, the child must 'know' or find out that a categorisation of words into parts of speech is relevant, find out what the part of speech tags for its language are, and find out for each word what its part of speech tag is. In addition, each of the three words under investigation here is multiply ambiguous, and many of the candidate modifiees are ambiguous.

For these reasons only a few aspects of first language acquisition are considered here and various simplifications and idealisations are assumed. For example, the analysis starts from an orthographic transcription, enriched with annotations for hesitations, filled pauses, retracings, etc.[6] The ambiguity of the words cannot be avoided, but the focus here is on only one meaning of the words under investigation, viz. the meaning *very*.

[4]The version of the corpus in PaQu contains approximately 1.9 million tokens.

[5]The annotations in the examples are CHAT-annotations as used in CHILDES corpora.

[6]Though these annotations are not always correct and surely not complete in the actual CHILDES corpora.

It is also assumed that children 'know' or have somehow found out that they should be 'looking for' grammatical dependencies, e.g. head-complement relations, modifier-modifiee relations etc., and that they are able to do so (though it is not obvious how they achieve this).

Since this paper investigates modifier-modifiee relations, I specifically make a number of assumptions on *modification*. Modification has two aspects: a *syntactic* aspect, and a *semantic* aspect. Syntactic modification specifies the syntactic structure(s) in which modifiers and modifiees can occur. I assume that there is an operation of modification *M* that applies to two elements X and Y and creates a syntactic modification structure. I assume that it yields a single configuration, formulated here in terms of the structures assumed in the treebanks used here: X syntactically modifies Y by the operation M(X,Y) = [mod/X, hd/Y] (order irrelevant), i.e. a node X with grammatical relation *mod* (modifier) modifies a node Y with grammatical relations *hd* (head) if they are siblings of the same node.[7]

A syntactic modification relation has a semantic pendant. The study of the semantics of modification is of course a research field in itself. However, for the purposes of this paper minimal assumptions suffice: if X is a syntactic modifier of Y, the corresponding semantic modification is built up compositionally on the basis of the meaning of X ($[\![X]\!]$), the meaning of Y ($[\![Y]\!]$) and the meaning of the syntactic modification operation ($[\![M]\!]$), i.e. $[\![M]\!]([\![X]\!], [\![Y]\!])$. This assumption will play a crucial role in section 7.

I will also assume that children 'know' or find out that syntactic selection restrictions of a modifier on a modifiee are specified in terms of syntactic category. The notation *mod A*, *mod V*, and *mod N* specifies the property of a word or phrase that it can modify an A, V or N, resp.

6 Treebank Query Results

In this section, the main results for the queries for the three words *heel*, *erg* and *zeer* as modifiers are presented as reported by (Odijk, 2015), as well as for the results of these queries in the Basilex corpus and the Lassy-Large Wikipedia part.

The corpora in which the queries have been carried out are characterised in Table 1.

Corpus	#utts (k)	#tokens (m)	modality	spontaneity	formality
LASSY-Small	65	1	written	prepared	formal
CGN	130	1	spoken	mixed	mixed
VanKampenJAC	61	0.3	spoken	spontaneous	informal
VanKampenCHI	47	0.15	spoken	spontaneous	informal
CHILDES Dutch	545	1.9	spoken	spontaneous	informal
Basilex		13.5	written	prepared	formal
Wikipedia	8707	145	written	prepared	very formal

Table 1: Corpora analysed in this study and their characteristics.

Each corpus has been characterised in terms of its size, i.e. its number of utterances (where available) and its number of tokens, its modality (written language or (transcripts of) spoken language), the spontaneity of its content and an indication of its formality. LASSY-Small is a treebank for written Dutch of app. one million tokens. Its written text is explicitly prepared and rather formal (e.g. it does not contain social media and usenet data). The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) treebank contains app. one million tokens for spoken Dutch. It has several subcomponents, differing in spontaneity and formality (e.g., it contains prepared read speeches but also spontaneous conversations). The Van Kampen corpus is one of the corpora in the CHILDES collection for Dutch. The child-directed utterances (VanKampenJAC) were investigated separately from the utterance of the target children (VanKampenCHI). The Van Kampen corpus contains transcriptions of the natural interaction between parents and children. I also investigated the whole CHILDES collection for Dutch.

The BasiLex corpus consists of 13.5 million tokens[8] of texts written for children in primary education.

[7] In order to properly work on the flat structures in the treebank which allow more than 2 siblings, the formulation should be generalised somewhat, but this is not essential for the purposes of this paper.

[8] 11.5 million if interpunction symbols are ignored.

It contains various genres, with 40% of the tokens coming from educative materials, 40% from child literature, and 20% from media (newsfeeds, subtitles, etc.). The time coverage is 1976-2013. At the time of these investigations, it was not possible yet to host a full treebank for the Basilex corpus. For that reason, a subcorpus was created by selecting all sentences containing any word form of the lemmas *heel*, *erg* or *zeer* and the resulting corpus was uploaded in PaQu. It is known there as the corpus HEZ-Basilex-JO, shared with everyone who is logged in.[9] This subcorpus contains 26,239 sentences. PaQu parsed these sentences using Alpino. Several of the sentences come from educational material, which often contains words in alternative spellings (e.g. *ver-schrik-ke-lijk* instead of *verschrikkelijk* 'horrible', *SLAAAP* instead of *slaap* 'sleep'), and exercises with incomplete words or lists of alternative words. In such cases, the automatic parses are often wrong:

(3) Zijn kinderen hebben 'm erg gemi...
his children have him very mis...
'His children mis... him very much'

(4) erg blij / bijt / bij
very glad / bite / by
'very glad / bite / by'

Finally, the Wikipedia part of Lassy-Large contains 145 million utterances of carefully prepared written language and it is, as an encyclopedia, very formal in nature. It is part of the (550 million token) SoNaR Corpus. Querying the whole treebank for the SoNaR Corpus with the treebank search applications is, despite the development of special techniques to speed up querying (Vandeghinste and Augustinus, 2014; Vanroy et al., 2017), unfortunately not yet possible for us, though the Institute for the Dutch Language recently made a version of GrETEL 4.0 available in which each of SoNaR's components can be searched separately.[10]

6.1 Mapping D-COI Part of Speech Tags

The treebanks consulted use the *de facto* standard for part of speech tagging of Modern Dutch words, so-called D-COI tags (Van Eynde, 2005). This tag set makes distinctions that differ somewhat from what is needed here. I describe here how I reclassified D-COI tags to the distinctions I want to make: Some tags map directly on tags I use, e.g. *adj* = adjective maps to *A*, *ww* = verb maps to *V*, *n* = noun maps to *N*, *vz* = adposition maps to *P*. For other tags the mappings are slightly more complex:

- *vnw* = pronoun. The pronominal nature of a word is an important morpho-syntactic distinction, but in my view it is independent of part of speech assignment. Words with the D-COI tag *vnw* were automatically mapped to *A* (e.g. for *veel* 'many', *weinig* 'few' and their comparative and superlative forms) or to *N* (e.g. for *wat* 'a few').
- *bw* = adverb. Words with D-COI tag *bw* are manually mapped to *A* or *P*, depending on the word.
- *mwu* = multiword unit. The characterisation of a word combination as a multiword unit is an important distinction, but in my view it is independent of part of speech assignment. Word combinations labeled with the D-COI tag *mwu* were manually mapped to *A*, *N*, *V* or *P* depending on the specific word combination.
- *tw* = numeral. Words with the D-COI tag *tw* are mapped to *N*.

The queries search for the words *heel*, *erg* or *zeer* when occurring as a modifier (grammatical relation *mod*). I specifically also searched for sentences that contain *heel*, *erg* or *zeer* and a predicative (*predc*) or locative (*ld*) complement, because such sentences are likely to contain incorrectly analysed examples of modification of adpositions. I also searched for uses of these words with a different grammatical relation: these should be irrelevant if the parse is correct but might contain misparsed examples.

[9] Everybody can log in by just using one's e-mail address.
[10] `https://portal.clarin.inl.nl/chn-gretel/ng/home`.

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	295.6	0.0	0.0
CGN	2899.4	0.0	7.9
VanKampenJAC	2191.1	3.3	3.3
VanKampenCHI	1616.9	0.0	6.5
CHILDES Dutch	2512.4	3.2	8.5
Basilex	172.0	0.0	1.7
Wikipedia	90.5	0.0	0.3

Table 2: Results of queries for *heel* as a modifier in a variety of corpora (relative frequency per million tokens).

6.2 Main Results for *heel*

Table 2 summarises the query results for *heel* as a modifier.

Some remarks on these figures are required. I will discuss some cases where *heel* appears to modify a verb (6.2.1) or an adposition (6.2.2).

6.2.1 *heel* Modifying Verbs

First, I discuss some special cases of *heel* modifying a verb. In the query results for Lassy-Small one does find the part of speech code for verb in the treebanks ('ww') as being modified by *heel*, but these are artifacts of the structure of the treebank, in which adjectives derived from participial verbs are categorised as verbs, as in (5):

(5) Examples of adjectives derived from participles, which are categorised as *ww* (verb) in the treebank:
   a. heel gecompliceerd ('very complicated')
   b. heel overtuigend ('very convincing')
   c. heel vervelend ('very boring/unpleasant')

Second, under the substantivised use of infinitives the word is also characterised as *ww*, though it has actually converted to a noun. The modifier *heel* only has the interpretations it has as a modifier of a noun ('whole') in such constructions. See (6):

(6) Examples of substantivised verbs that are categorised as *ww* (verb) in the treebank:
   a. het hele ... gebeuren ('the whole ... happening')
   b. hun hele hebben en houden
      their whole have and hold
      'all their possessions'

In the Spoken Dutch Corpus, there are some examples of *heel* modifying verbs, but they are ill-formed for me, and are used almost exclusively by Flemish speakers in informal registers. I found similar examples in the SoNaR corpus. I suspect that people who use this can use *heel* in the sense of *geheel* 'completely' (and this is how I glossed them in (7)). This surely requires further investigation, but I will not deal with these examples here. Some examples:

(7) *heel* modifying verbs by Flemish speakers in informal registers:
   a. ...heel te verdwalen... ('to get completely lost')
   b. ...heel omgebouwd... ('completely rebuilt')

In VanKampenJAC also one example occurs (session laura030.cha, speaker JAC):

(8) Ik kijk heel uit
   I look very out
   'I am very cautious'

The example is ill-formed for me, and in this particular case I could check the example with the speaker. She confirmed that the sentence is ill-formed for her too, and that it must have been a performance or transcription error. Such an example, and several other examples, do show that ill-formed input is offered to children, who must thus be robust against such ill-formed input.

In the Dutch CHILDES as a whole more examples of *heel* modifying a verb occur, but these are all utterances by children, who apparently did not get the rules yet. Researching them is outside the scope of this paper.

6.2.2 *heel* Modifying Adpositions

There are also some cases where *heel* modifies or appears to modify an adposition. First, there are some examples in which *heel* modifies an adverbial PP:

(9) a. heel in de verte
very in the far-th
'at a very great distance'

b. heel in het begin
very in the beginning
'in the very beginning'

c. heel af en toe
very off and to
'very infrequently'

d. heel in de verte
very in the distance
'at a very great distance'

I found ten different cases (the four of (9) and *heel in het algemeen* lit. very in the general ('very generally'), *heel in het bijzonder* lit. very in the particular 'more particularly', *heel in het kort*, lit. very in the short 'very briefly', *heel op het laatst* lit. very at the last 'at the very end', *heel uit de verte* lit. very from the far-th 'from a very great distance', and *heel aan het eind* lit. very at the end 'at the very end'.) Such examples were found in most corpora (CGN, VanKampenJAC, CHILDES Dutch, Basilex and Wikipedia).

Furthermore, I found one additional example:

(10) ...'t heel voor de hand ligt...
...it very before the hand lies...
'...it is very obvious...'

Though the present participle form of this expression *voor de hand liggend* is adjectival in nature and is often modified by *heel*, modification of the verbal form is ill-formed according to my judgement as a native speaker. I will assume it is a performance error.

Finally, I found one example (by a Flemish speaker) where *heel* modifies an adposition and where it probably means 'completely': *heel beneden* 'completely downstairs'.

In CHILDES, there are several examples of *heel* modifying an adposition in the children's speech but also one by an adult (which is ill-formed, according to my judgement as a native speaker):

(11) heel iets naar buiten BOU mat20501.429 (father)
very somewhat to outside
'a little bit to the outside (?)'

again showing that the language acquisition device must be robust against ill-formed input.

6.3 Main Results for *erg*

Table 3 shows the treebank query results for modification by *erg*.

There is one example in the children's speech in VanKampenCHI where *erg* appears to modify an adposition, but no other peculiarities.

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	156.0	13.7	5.5
CGN	324.6	78.1	13.2
VanKampenJAC	112.5	49.6	0.0
VanKampenCHI	77.6	6.5	6.5
CHILDES Dutch	189.7	44.1	2.1
Basilex	324.5	73.8	3.5
Wikipedia	128.3	12.2	2.1

Table 3: Results of queries for *erg* as a modifier in a variety of corpora (relative frequency per million tokens).

6.4 Main Results for *zeer*

Table 4 shows the treebank query results for modification by *zeer*.

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	307.4	7.3	2.7
CGN	207.0	7.9	1.8
VanKampenJAC	6.6	6.6	0.0
VanKampenCHI	6.5	0.0	0.0
CHILDES Dutch	6.4	2.7	1.6
Basilex	26.7	1.7	0.3
Wikipedia	342.0	18.3	1.9

Table 4: Results of queries for *zeer* as a modifier in a variety of corpora (relative frequency per million tokens).

There are a few examples that might involve modification of an adposition by *zeer*, but they might also involve modification of the verb or the whole verb phrase. It concerns modification of the expression *op prijs stellen* lit. *on price put* 'appreciate'[11] and of the expression *in de smaak vallen* lit. *in the taste fall* 'like (with arguments reversed)'[12], all by adults. They are analysed in the treebank as modifying the verb and that is surely defensible and actually most likely the correct analysis.

6.5 Summary of the Query Results

The results for all corpora except Basilex and Wikipedia were already reported in (Odijk, 2015) and (Odijk, 2016). His findings for the child-directed speech in these corpora can be summarised as follows:

- Of the three words *heel*, *erg* and *zeer*, *heel* occurs most frequently.
- There is an overwhelming number of cases where *heel* modifies an adjectival phrase (>92%).
- Modification of verbal phrases by *heel* does not occur.
- There are many examples where *erg* modifies an adjectival phrase, but also a significant number of cases where it modifies a verb phrase.
- There are very few examples of *zeer* modifying an adjectival phrase, and also very few in which it modifies a verb.
- There are no clear examples with *erg* or *zeer* modifying a PP.

For the problem under investigation, this means:

[11] Utterances jos20021.354 and tom20507.71 from the Groningen Corpus.
[12] Utterance iri30323.1283 from the Groningen corpus.

Corpus	*heel*	*erg*	*zeer*
Basilex	1.7	3.5	0.3
Wikipedia	0.3	2.1	1.9

Table 5: Relative frequency per million tokens of *heel*, *erg* and *zeer* modifying an adposition in Basilex and Wikipedia.

- The data seem appropriate for acquiring the property that *heel* modifies adjectival but no verbal phrases.
- It is less clear how modification of PPs can be excluded, since there are some examples where *heel* modifies PPs.
- The absence of data for *zeer* makes it difficult to state anything about the acquisition of its modification potential.

In order to address the latter two problems, more data are needed. Unfortunately, there are no other CHILDES data for Dutch that are relevant in this context. However, Odijk (2016) observes that *heel* occurs very early in the children's speech (1;11), with *erg* occurring only a year later (2;10), and *zeer* very late (4;8). He ascribes the late occurrence of *zeer* to its more formal character. A corpus of data typical for the input that children hear or read from the age of 5 years old would be ideal to address these problems. The BasiLex corpus (Tellings et al., 2014) is exactly such a corpus: it contains texts that are directed at children at primary school. In addition, it is significantly larger than the CHILDES corpora. Because of the late acquisition of *zeer*, BasiLex's focus on texts that are targeted at children between the ages of 6 and 12 appears to make it particularly appropriate for investigating the modification potential of *zeer*.

I used PaQu to investigate the properties of modifiees of *heel*, *erg* and *zeer*, respectively. A manual analysis of the query results was carried out in order to map the more refined distinctions made by PaQu onto the distinctions needed here, and to correct wrong parses by Alpino.

The crucial data are presented together in Table 5. Strikingly, examples with *heel* modifying a PP are more frequent (1.7 / million tokens) than *zeer* modifying a PP (0.3 / million tokens) in Basilex, but this does not have the effect that the adult grammar allows modification of PPs by *heel* in general. Conversely, despite their low frequency even in this larger corpus, the adult grammar allows modification of PPs by *zeer* generally. In addition, the frequency of *zeer* modifying PPs is so low, that one might wonder whether they are taken into account at all in the acquisition process. After all, utterances may be analysed incorrectly by the child, or might be misheard, or might be mispronounced by the speaker, so it seems reasonable to require a minimum number of occurrences of a phenomenon before it is taken into account in adapting lexical properties or grammar rules, at least in the case of unconscious acquisition, as is the case here. In any case, the language acquisition procedure must be robust against some noise ((Yang, 2016, 13) and references there).

Concluding, despite a larger and more representative corpus, the same questions still lie before us:

- Why does the presence of PPs modified by *heel* not lead to generalising the modification potential of *heel* to PPs generally?
- Why is the modification potential of *zeer* generalised to PPs generally despite its very low frequency?

Perhaps also the Basilex corpus is not big enough to get a representative overview. Therefore, an even larger corpus, the Wikipedia part of Lassy-Large (145 million tokens), has been investigated. Though this corpus is not representative for language acquisition at all, it might give us insight into the degree of representativity of the CHILDES corpora and the BASILEX corpora for the problem at hand.

It is clear that *zeer* occurs much more often here than in the earlier corpora as a modifier of adjectival, verbal and adpositional phrases. However, even here, in this large and very formal corpus, the frequency

of *zeer* modifying a PP is extremely low (1.9 per million). Examples of *heel* modifying a PP are less frequent in this corpus, but I ascribe this to the rather formal nature of this corpus. I conclude that even this very large and very formal corpus does not provide an answer to the major questions that the data raise.

7 Towards Analysis of the Data

In this section, a tentative attempt to analyse the data is presented. I will first discuss the possibility of analysing these data using Yang's theory on the *Sufficiency Principle* in section 7.1. I argue that this theory does not contribute to explaining these data. In section 7.2 I argue that the combinations of *heel* modifying adpositions are idiosyncratic in nature and cannot provide evidence for a productive rule of modification. Finally, in section 7.3 I sketch my proposal for the acquisition of these constructions.

7.1 The Sufficiency Principle

Since the modifier *zeer* has properties based on very little data, it is natural to investigate whether it has these properties not from direct positive evidence but from a productive rule that applies to it. It seems to me that there is no productive rule in Dutch that determines the modification potential of degree modifiers, so if this assessment of the facts is correct, it is unlikely that any theory of productivity of rules will contribute to addressing this problem. If there would be a productive rule, it should be a rule that predicts that degree modifiers can modify adpositions if it is to account for the fact that *zeer* has the potential to modify adpositions despite very little positive evidence for this.

One approach that addresses the issue of the productivity of rules has been proposed by (Yang, 2016). He considers (inter alia) the exclusively predicative use of *A*-adjectives (e.g. *asleep*, *awake*, *alone*, *away*) and dative alternation in English. He claims that the relevant words belong to a class, and that the members in this class generalise their modification or complementation potential in accordance with what he calls the *Sufficiency Principle* (Yang, 2016, 177):[13]

(12) Let *R* be a generalisation over *N* items, of which *M* items are attested to follow *R*. *R* can be extended to all *N* items if and only if: $N - M < \theta_N$ where $\theta_N = N/\ln N$.

Applying this hypothesis to the problem of this paper requires first of all establishing a class that the degree modifiers belong to. They certainly do not have morphological properties in common, but one might consider them as members of the semantically defined class of degree modifiers. Yang defines the class of verbs to account for dative alternation phenomena also semantically (as 'verbs of caused possession that involve the transfer of objects, entities or abstract information' (Yang, 2016, 201)). Second, a rule that applies to this class must be postulated. Actually, the relevant rule should predict that *mod P* is a property of degree modifiers. However, if this rule is productive, it will be impossible to have exceptions to this rule that do not have *mod P* as a property (such as *heel*) if negative evidence plays no role in first language acquisition (as is generally assumed): we basically then have an instance of Baker's Paradox here (Baker, 1979). I inventoried around 145 words and expressions that can act as degree modifiers.[14] The largest subclass, members of which can have the property *mod A* | *mod V* | *mod P*, contains minimally 35 and maximally 83 elements.[15] Even if all 83 belong to this class, applying the *Sufficiency Principle* (12) yields the following result: N = 145, and, for the postulated rule, M = 83: $145 - 83 = 62$. This should be smaller than θ_{145} but it is larger than θ_{145}, where $\theta_{145} = 145/ln145 \approx 29$. I conclude that even the best candidate rule under these assumptions is predicted by the *Sufficiency Principle* not to be productive. I conclude that the *Sufficiency Principle* cannot account for the relevant facts.[16]

[13] Yang actually has 'if and only **iff**', which I assume is a typo; He also has := instead of the equal sign, of which I also assume it is a typo.

[14] I did so by crucially using a different application developed in the context of CLARIN: Cornetto, which offers a search interface to the Dutch WordNet (Vossen et al., 2013).

[15] I was not yet able to determine the property *mod P* for all these words: my intuitive judgements on theses examples are uncertain and I was not yet able to do corpus searches for all these words. Fortunately, this is not crucial here, as will become clear below.

[16] But of course, this does not mean that the *Sufficiency Principle* is wrong.

It might be investigated what predictions the *Sufficiency Principle* makes *during* first language acquisition, to model various stages of the language acquisition process, but I will leave that to future research.

7.2 The Idiosyncratic Nature of *heel* Modifying Adpositions

In this section, I argue that constructions in which *heel* modifies adpositions are idiosyncratic constructions that must be acquired one by one and that do not constitute evidence for a productive rule such as the modification rule.

The first consideration in this regard comes from a closer look at the PPs modified by *heel*, e.g. *heel in de verte*. I stated that *heel* syntactically modifies the adpositional phrase (PP) *in de verte*. However, this PP expresses a location, and locations cannot be semantically modified by degree modifiers such as *heel*, *erg* and *zeer*. This is clear from the examples in (13) and (14):

(13) hij staat (*erg) op het veld
he stands very on the field
'He is standing (*very much) on the field'

(14) a. Zij zit in de put
She sits in the well
'She is sitting in the well / She is depressed'

b. Zij zit erg in de put
She sits very in the well
'*She is sitting very much in the well / She is very depressed'

Modifying the location expressed by the PP *op het veld* 'on the field' by the degree modifier *erg* leads to ill-formedness (13). The phrase *in de put zitten* in (14) is ambiguous between a literal interpretation (with the PP *in de put* as a location 'in the well') and an idiomatic interpretation (in which *in de put* expresses a mental state 'depressed'). Modifying the PP by a degree modifier disambiguates the PP, which then only has the mental state interpretation.

If degree modifiers cannot semantically modify locations, then what does *heel* modify semantically in *heel in de verte*? A look at the gloss and the translation makes this clear. *Heel* in this expression semantically modifies the adjective *ver* which is part of a derived noun (*ver-te* i.e. *far-th*) inside a noun phrase contained in the PP syntactically modified by *heel*, cf. the translation *at a **very great** distance*. This meaning cannot arise from the normal rule of modification with compositional semantics (see section 5), which requires that the meaning of the full expression is derived from the meaning of *heel* and the meaning of the whole PP *in de verte*. I thus conclude that these constructions cannot be seen as special instances of the normal rule of modification. In fact, the semantic modification of the morpheme *ver-* by *heel* cannot be part of any productive linguistic rule, and the expression must thus be stored as an instance of an idiosyncratic mapping between form and meaning. This is confirmed by the fact that only a handful of different examples of this construction were found (in quite large corpora) and by the fact that no or only very limited variation is possible. For example, in example (9d), one cannot replace the noun by semantically related nouns (15):

(15) a. * heel in de nabijheid
very in the closeness
'at a very small distance'

b. * heel in de hoogte
very in the height
'at a very great height'

c. * heel in de diepte
very in the depth
'at a very great depth'

and for *de verte* only the prepositions *in* 'in' and *uit* 'out of' are possible, but e.g. *naar* 'towards' is not (16):

(16) * heel naar de verte
very to the distance
'towards the far distance'

Similar restrictions hold for the other examples. This thus disqualifies these examples as instances of regular modification of a PP by *heel*.

This analysis, however, does not apply to the expression *heel af en toe*. Here the part *af en toe* has both an unusual syntax (coordination of two adpositions) and an idiosyncratic meaning ('occasionally') but *heel* modifies the part *af en toe* as a whole, in accordance with the rule for modification. I still argue that the combination of *heel* and *af en toe* is idiosyncratic.[17] I observe that *af en toe* can**not** be modified by *erg* and *zeer*.[18] In addition, the expression *nu en dan* lit. now and then 'occasionally', which is a synonym or near-synonym of *af en toe*, cannot be modified by any of the words *heel*, *erg* or *zeer*.[19] I therefore conclude that the combination *heel af en toe* is also an idiosyncratic combination and not an instance of the regular modification rule.

7.3 Towards an Analysis

If the cases where *heel* modifies adpositions are idiosyncratic and do not provide evidence for general rules or principles, a simple statistical learning strategy can account for the data. I make several assumptions: (1) that the modification potential of words is acquired by positive evidence only; (2) that each property of a lexical item has an associated activation score, which increases each time there is evidence in the input for this property; (3) that the activation score must be higher than a threshold θ_{min}. This is necessary to be robust against ill-formed, misheard or mis-analysed input. It then follows that *heel* selects A, and only A (no positive evidence for mod V or mod P), while *erg* and *zeer* select not only A but also V and P.

The question remains what the value of θ_{min} is or how it is determined. This is a matter that has to be determined empirically by studying multiple cases. No firm conclusions can be drawn on the basis of the phenomenon studied here alone. It is possible that assuming a decay function, which lowers the activation score over time, might play a role here too.[20] Here I speculate that θ_{min} must be very low to account for *zeer* selecting P (< 0.3 / million tokens), and it might also be a function of the number of relevant examples encountered, so that its value actually increases over time if there is sufficient input. Future research will have to clarify whether these speculations correspond to the facts.

8 Concluding Remarks and Future Work

This paper analysed data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning 'very'. The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular treebank query applications. The use of treebanks was necessary because of the high ambiguity of the words. In addition, the use of the CLARIN applications made it possible to base the analysis provided in this paper on a far larger empirical base than would have been possible without these applications, and the applications enable the researcher to query the data efficiently.

[17] It was also suggested to me that *af en toe* might actually be an adjective instead of an adposition. It is not so easy to determine what category *af en toe* belongs to. Many standard tests are inconclusive. However, the so-called PP-over-V test (Broekhuis, 2013, 8) suggests that *af en toe* is adpositional, cf. the well-formedness of e.g. *je zou het bijna vergeten af en toe*, lit. one would it almost forget occasionally, 'one would occasionally almost forget it', with *af en toe* to the right of the infinitive *vergeten*. Whatever category it is, a different category assignment would not account for the idiosyncracies observed in the main text.

[18] A search in the 550 million token SoNaR Corpus, which contains 32,119 occurrences of *af en toe* yields exactly one result of *erg* modifying *af en toe*.

[19] The 550 million token SoNaR corpus contains 3,909 occurrences of *nu en dan*, and there are no occurrences of modification by *heel*, *erg* or *zeer*. There are 5 occurrences of the combination *zo heel* modifying *nu en dan*, but *zo* is obligatorily present in these constructions. I assume that *zo heel* is also an idiosyncratic combination.

[20] Such a decay function might provide an account of the phenomenon of language attrition.

References

Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. 2012. Example-based treebank querying. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

C.L. Baker. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4):533–581.

Robert Berwick. 1985. *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

Hans Broekhuis. 2013. *Syntax of Dutch: Adpositions and Adpositional Phrases*. Amsterdam University Press. http://www.oapen.org/record/462289.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 23, pages 281–297. Ubiquity, London, UK. DOI: http://dx.doi.org/10.5334/bbi.23. License: CC-BY 4.0.

Jan Odijk, Martijn van der Klis, and Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 46–55, Prague, Czech Republic, January 23-24. http://aclweb.org/anthology/W/W17/W17-7608.pdf.

Jan Odijk. 2015. Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14, December.

Jan Odijk. 2016. A Use case for Linguistic Research on Dutch with CLARIN. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015, October 14-16, 2015, Wroclaw, Poland*, number 123 in Linköping Electronic Conference Proceedings, pages 45–61, Linköping, Sweden. CLARIN, Linköping University Electronic Press. http://www.ep.liu.se/ecp/article.asp?issue=123&article=004, http://dspace.library.uu.nl/handle/1874/339492.

Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat, and Harald Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*. ELRA.

Steven Pinker. 1989. *Learnability and Cognition*. MIT Press, Cambridge, MA.

Agnes Tellings, Micha Hulsbosch, Anne Vermeer, and Antal van den Bosch. 2014. BasiLex: an 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208, 12/2014.

Frank Van Eynde. 2005. Part of speech tagging en lemmatisering van het D-COI corpus. CGN report, Centrum voor Computerlinguïstiek, KU Leuven, Leuven, Belgium, July. http://www.ccl.kuleuven.be/Papers/DCOIpos.pdf.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

Vincent Vandeghinste and Liesbeth Augustinus. 2014. Making large treebanks searchable. The SoNaR case. In *Proceedings of the LREC 2014 2nd workshop on Challenges in the Management of Large Corpora (CMLC-2)*, pages 15–20, Reykjavik. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-CMLC2%20Proceedings-rev2.pdf.

Bram Vanroy, Vincent Vandeghinste, and Liesbeth Augustinus. 2017. Querying large treebanks: Benchmarking GrETEL indexing. *Computational Linguistics in the Netherlands Journal*, 7:145–166, 12/2017.

Piek Vossen, Isa Maks, Roxanne Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: a lexical semantic database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch, Results by the STEVIN-programme*, Theory and Applications of Natural Language Processing, chapter 10, pages 165–184. Springer, Berlin Heidelberg.

Charles Yang. 2016. *The Price of Productivity: How Children Learn to Break the Rules of Language*. MIT Press, Cambridge, Mass.