Computational Linguistics in the Netherlands Journal 5 (2015) 3-14 Submitted 06/2015; Published 11/2015

Linguistic Research with PaQu

Jan Odijk J.ODIJK@UU.NL

UiL-OTS, Utrecht, Netherlands

Abstract

In this paper I illustrate the use of the PaQu (**Pa**rse and **Qu**ery) application for carrying out linguistic research. The major findings of this paper are: (1) PaQu is very useful for aiding researchers in efficient manual verification of hypotheses; (2) PaQu can even be used for automatic verification of hypotheses, provided some care is exercised; (3) the Dutch CHILDES data are too small to address certain research questions; and (4) the data found suggest several new hypotheses on the acquisition of lexical properties that should be further explored.

1. Introduction

In this paper I illustrate the use of the PaQu (**Pa**rse and **Qu**ery) application for carrying out linguistic research. I describe the results of a small experiment in linguistic research into first language acquisition. The experiment was carried out mainly to test the functionality of the PaQu application, but I selected an example related to a linguistic problem that fascinates me and that I have used earlier to guide developments in the CLARIN infrastructure.

The major findings of this paper are: (1) PaQu is very useful for aiding researchers in efficient manual verification of hypotheses; (2) PaQu can even be used for automatic verification of hypotheses, provided some care is exercised; (3) the Dutch CHILDES data are too small to address certain research questions; and (4) the data found suggest several new hypotheses on the acquisition of lexical properties that should be further explored.

In section 2 I sketch the background of this work. I introduce the basic facts to be investigated in section 3, make an assessment of these facts in section 4, and list a few of the many research questions that these facts raise in section 5. The data needed to address these research questions are introduced in section 6. These data are not rich enough for addressing the research questions directly, which is where PaQu starts playing a role (section 7). Section 8 evaluates the quality of the relevant parts of the parses generated by Alpino in PaQu for the adult speech in the Dutch CHILDES utterances that are relevant to the current research. It does the same for the children's utterances. Section 9 analyzes the results for all relevant utterances by adults. Section 10 summarizes the conclusions and section 11 describes future work that can and must be done to address the research questions.

2. Background

CLARIN-NL has made available a wide range of web applications for search in and analysis of linguistically annotated corpora for Dutch. These include OpenSONAR, FESLI, COAVA, and MIMORE. CLARIN Flanders made available GrETEL.[1] However, each of these interfaces applies to a specific set of text corpora only. Many linguists want to have such search and analysis opportunities also for the text corpora they are currently working with. To that end, the CLARIN-NL project commissioned the development of *PaQu* and *AutoSearch*.[2]

[1] For more information on these (and other) web applications and links to them, see the CLARIN-NL portal faceted search for resources (`http://portal.clarin.nl/clarin-resource-list-fs`).

[2] Adding upload facilities to GrETEL was also requested but its development could not be started at the time. It is being implemented currently, though (see section 11).

PaQu is a web application developed by the University of Groningen. It enables one to upload a Dutch text corpus. This text corpus is either already parsed by the syntactic parser for Dutch Alpino[3] (van der Beek et al. 2002), or if not, PaQu can have it automatically parsed by Alpino. After this, it is available in the word relations search interface of PaQu, as well as via PaQu's XPATH interface.

AutoSearch[4] is a web application developed by INL. Here FoLiA or TEI formatted Dutch text corpora containing (extended) PoS codes (e.g. as created by the Frog[5] (van den Bosch et al. 2007) part of speech tagger in TTNWW[6]) can be uploaded and searched via a Corpus of Contemporary Dutch[7]-like search interface. This application will not be discussed in this paper any further.

Both applications are available in the CLARIN infrastructure and can be accessed via the CLARIN-NL portal[8].

3. Basic facts

In this section I introduce the basic facts related to the problem that I am interested in investigating. It is a specific case of the problem of the acquisition of lexical properties by first language learners.

The three words *heel*, *erg* and *zeer* are (near-)synonyms meaning 'very', i.e. (stated informally) they modify a word 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) predicates only, while *erg* and *zeer* can modify not only adjectival, but also verbal (V) and adpositional (P) predicates. This is illustrated in example (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 predicate *blij* 'glad' can be modified by each of the three words. In (1b) the (idiomatic) prepositional predicate *in zijn sas* can be modified by *zeer* and *erg* but not by *heel*.[9] The same holds in (1c) for the verbal predicate *verbaast*.[10] In English, something similar holds for the word *very*: it can only modify adjectival predicates. For verbal and prepositional predicates one cannot use *very* but one can use the expression *very much* instead:

(2) a. He is very happy about it

[3] `http://www.let.rug.nl/vannoord/alp/Alpino/`

[4] `http://portal.clarin.nl/node/4222`

[5] `http://ilk.uvt.nl/frog/`

[6] `http://portal.clarin.nl/node/1964`

[7] `http://corpushedendaagsnederlands.inl.nl/`

[8] `http://portal.clarin.nl/`

[9] One reviewer suggested to include the word *geheel* in the analysis, since this word can modify prepositional phrases and there might be complementary distribution between *heel* and *geheel*. However, *geheel* has a different meaning ('completely'), so it is not obvious that it is relevant in this context. What is relevant is that (Odijk 2014, 27) found, using the OpenSONAR application, that some people use *heel* instead of *geheel* as a modifier of prepositional phrases (which is completely out for me). Since this also involves a different meaning of *heel*, I will not deal with it in this paper.

[10] Or maybe the whole VP *verbaast mij*.

b. He is very *(much) in love with her
c. It surprised me very *(much)

There is a lot more to say about these data, and there are more relevant data to consider and some qualifications to be made. I refer to (Odijk 2011) and (Odijk 2014) for details.

4. Assessment of the facts

The distinctions I 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 difference can be derived from semantic properties.[11] 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 in any of the differences that I am aware of.

The first difference concerns meaning. It is plausible that *erg* has a meaning that differs slightly from *heel* and *zeer*[12] But this opposes *erg* to {*heel*, *zeer*} instead of *heel* to {*erg*, *zeer*}.

The second difference relates to adverbial inflection. The words *heel* and *erg* can be inflected when they modify an inflected adjective:[13]

(3) a. Hij heeft heel / erg / zeer grote handen
he has very / very / very big hands
'He has very big hands'

b. Hij heeft hele / erge / *zere grote handen
he has very-E / very-E / very-E big hands
'He has very big hands'

This is not part of the standard language but occurs a lot in informal and spoken language. The impossibility of *zere* in (3b) cannot be ascribed to morphological restrictions, since this very form is well-formed as an inflected form of the adjective *zeer* in the reading *painful*.[14] This difference, however, opposes {*heel*, *erg*} to *zeer*, and not *heel* to {*erg*, *zeer*}, as would be required.

A third difference relates to comparative and superlative forms. The word *erg* can form comparative and superlative forms when it modifies predicates. That is completely impossible for *heel* and *zeer*:[15]

(4) a. Jan is *heler / erger / *zeerder ziek (dan Piet)
Jan is very-ER / very-ER / very-ER ill (than Piet)
'Jan is more awfully ill than Piet'

b. Jan is het *heelst / ergst / *zeerst ziek
Jan is the very-EST / very-EST / very-EST ill

[11] See (Odijk 2011) for more examples supporting this conclusion.

[12] In particular, it seems to have an additional meaning aspect that evaluates the high degree as negative, close to English *awfully*.

[13] I use *E* as a code for the inflectional suffix on the adjective.

[14] And if it could be derived from morphological considerations, the example would lose its relevance here completely.

[15] *ER* is used as a code for the comparative suffix; *EST* as a code for the superlative suffix.

'Jan is most awfully ill'

But this opposes *erg* to {*heel*, *zeer*} instead of *heel* to {*erg*, *zeer*}.

A fourth difference concerns modification: *erg* can be modified itself, while { *heel*, *zeer*} cannot:[16]

(5) a. Jan is #heel heel / heel erg / *heel zeer ziek
Jan is very very / very very / very very ill
'Jan is very seriously ill'

b. Jan is *erg heel / ?erg erg / *erg zeer ziek
Jan is very very / very very / very very ill
'Jan is very seriously ill'

c. Jan is *zeer heel / zeer erg / #zeer zeer ziek
Jan is very very / very very / very very ill
'Jan is very seriously ill'

Examples containing two identical modifiers are marked with # because they are well-formed under a parse that is irrelevant here, viz. as repeated degree modifiers of the adjective *ziek* (cf. English *very very ill*). I ignore this reading here. This difference opposes *erg* to {*heel*, *zeer*} instead of *heel* to {*erg*, *zeer*}.[17]

A fifth difference concerns pragmatic properties of these words: *zeer* is quite formal, while *heel* and *erg* are pragmatically neutral. But this opposes *zeer* to {*heel*, *erg*} instead of *heel* to {*erg*, *zeer*}.

Though a full account of the differences observed here is beyond the scope of the current paper, it can be safely concluded that none of the observed differences opposes *heel* to {*erg*, *zeer*}, so that it is very unlikely that the differences found here can be related to the differences in modification potential.

I conclude that the differences in modification potential of the words *heel*, *erg* and *zeer* cannot be derived from other facts and must be acquired by learners of Dutch.

5. Research questions

The simple facts described in the previous sections raise many research question related to language acquisition. Examples of these research questions are:

(6) a. How can children acquire the fact that *erg* and *zeer* can modify A, V and P predicates (in L1 acquisition)?

b. How can children acquire the fact that *heel* can modify A but **canNOT** modify V and P predicates (in L1 acquisition)?

c. What kind of evidence do children have access to for acquiring such properties?

d. Is there a relation with the time of acquisition?

e. Is there a role for indirect negative evidence (absence of evidence interpreted as evidence for absence)?

[16] The example with *erg* modifying *erg* is according to my judgment well-formed, though it is stylistically infelicitous, and the (irrelevant) repeated modifier reading is much more prominent.

[17] The word *zeer* appears to allow the modifier *te* 'too', but I assume that this is an unproductive fixed combination. All three words appear to allow *zo* 'so' as a modifier, though perhaps *zo* actually modifies the whole containing phrase rather than just the modifier.

Obviously, this paper cannot address all these questions. The main purpose of this paper is to show that, by using PaQu, research questions such as the ones in (6) can be addressed in a better and more efficient manner than without this (or a similar) tool. In this paper, I will focus on research question (6c)

In order to address these research questions, data are needed that can provide evidence for these questions. Fortunately, such data exist, and the Dutch CHILDES corpora are the most important ones in this context.

6. Dutch CHILDES corpora

The Dutch CHILDES corpora[18] are accessible via the CLARIN Virtual Language Observatory (VLO)[19] or directly via Talkbank[20] and contain relevant material to investigate the research questions formulated in section 5. They contain transcriptions of dialogues between children acquiring Dutch on the one hand, and adults (mostly parents) and in some cases other children on the other hand, and a lot of additional information about the context, setting, age of the child, etc.

However, a serious problem for the investigation is that the words being investigated are, as any decent word in natural language, highly ambiguous. Table 1 describes the ambiguity. For example, the word *heel* is 6-fold ambiguous. This ambiguity is partly solved by taking into account morphosyntactic and syntactic factors. For *heel* as a finite verb (Vf) the ambiguity reduces to 2, which cannot be further resolved by morphosyntax or syntax: 'heal' and 'receive' (of stolen goods). As an adjective (A) *heel* is 4-fold ambiguous. The ambiguity is partially resolved by taking into account its syntactic properties with regard to modification: if it modifies an adjective (mod A), the ambiguity is resolved to the single meaning 'very'; if it modifies a noun (mod N), the ambiguity is reduced to 3: 'whole', 'in one piece' or 'large'. If it is used as a predicative complement, it can only mean 'in one piece'.[21]

The Dutch CHILDES corpora do not contain any information about the meanings of its word occurrences. However, as is clear from Table 1, most of the ambiguities can be resolved by taking into account morpho-syntactic and syntactic properties of the word occurrences. Unfortunately, the Dutch CHILDES corpora do NOT have (reliable) morpho-syntactic information (part of speech tags) and they do not contain syntactic information for the utterances at all.

For this reason, (Odijk 2014, 91) carried out a manual analysis in terms of morpho-syntactic and syntactic information to disambiguate the occurrences of *heel*, *erg* en *zeer* in adult utterances of the Dutch CHILDES Van Kampen subcorpus. With PaQu, however, one can largely automate the disambiguation process, and I carried out a small test to investigate how well this works.

7. PaQu

PaQu[22] is a web application developed by the University of Groningen. It enables one to upload a Dutch text corpus. This text corpus is either already parsed by Alpino, or if not, PaQu can have it automatically parsed by Alpino. After this, it is available in the word relations search interface of PaQu (an extension of the Groningen Word Relations Search application[23] originally developed by (Tjong Kim Sang et al. 2010), as well as via PaQu's XPATH interface.

[18] I considered the subcorpora DeHouwer, Gillis, Groningen, Schaerlaekens, VanKampen, Wijnen and Zink, but not CLPF.

[19] `http://catalog.clarin.eu/vlo/search?fq=languageCode:code:nld&fq=collection:TalkBank`

[20] `http://childes.talkbank.org/data/Germanic/Dutch/`

[21] I use the following notation in the table: *Mod X* means that the word can modify a word of category X; *Mod X Y Z* means that a word can modify words of any of the categories X, Y, or Z; *predc* stands for *can occur as predicative complement*; Dutch distinguishes two values for gender: *uter* (i.e., not neuter) and *neuter.Vf* stands for *finite verb form.*

[22] `http://portal.clarin.nl/node/4182`

[23] `http://www.let.rug.nl/~alfa/lassy/bin/lassy`

Word	Morphosyntax	Syntax	Meaning
heel	A	Mod N	1. ‘whole’ 2. ‘in one piece’ 3. ‘large’
		predc	‘in one piece’
		Mod A	‘very’
	Vf		1. ‘heal’ 2. ‘receive’
erg	N	uter	‘erg’
		neuter	‘evil’
	A	Mod N, predc	‘bad’, ‘awful’
		Mod A V P	‘very’
zeer	N		‘pain’
	A	Mod N, predc	‘painful’
		Mod A V P	‘very’

Table 1: Ambiguity of the words *heel*, *erg* and *zeer*

For the specific problem dealt with here, we need, for each of the words *heel*, *zeer* en *erg*, a characterisation of the part of speech of the head word it is a dependent of and the label of the dependency relation (grammatical relation) holding between them. PaQu offers a dedicated interface precisely for this. The relevant queries are not easily expressed in XPATH[24], which makes GrETEL (after it has been extended with corpus upload facilities) less suited for this particular problem (but it might be more suited for other problems). The *AutoSearch* application is not suited for this problem either since it supports annotations at the token level only.

The output of PaQu is a list of utterances that match the query, and (partially user-definable) statistics on properties of matched words and matched triples of the form (property of dependent word, grammatical relation, property of head word).[25] Each of the matches and each of the statistical aggregates contains links with automatically generated queries for exploring specific subcases in more detail.

PaQu accepts as input plain text (in multiple varieties) or a text corpus parsed by Alpino in the LASSY XML[26] format. It currently does not allow a CHILDES corpus (in CHAT format (MacWhinney 2015)) directly as input.

In order to evaluate the quality of the automatically generated parses for this research, I selected all transcriptions of adult utterances in the Van Kampen corpus that contain one of the words *heel*, *erg*, or *zeer*. The transcriptions contain all kinds of mark-up that Alpino cannot deal with, so I

[24] Such a query has to take into account not only headed structures but also coordinated structures and co-indexed nodes in the syntactic structure. In addition, the dependent word can be contained in a phrase that is a dependent of the head word.

[25] Where *properties* include *word form*, *lemma*, and *part of speech.*

[26] `http://www.let.rug.nl/vannoord/Lassy/alpino_ds.dtd`

wrote a script to clean out this mark-up.[27] In most cases the mark-up can simply be left out, but in other cases the content of the mark-up is preferably retained. Example (7) illustrates both cases:

(7) a. ja , maar <we be> [//] we bewaren (he)t ook
yes , but <we kee> [//] we keep (i)t too
'yes, but we keep it too'
b. ja , maar we bewaren het ook

The mark-up [//] indicates that the preceding text between angled brackets is a false start ('retracing', (MacWhinney 2015, 73)). The cleaning script removes the mark-up and the false start text. The mark-up *(he)t* is the CHAT parenthesis notation for the 'abbreviated' form *'t* (MacWhinney 2015, 53). The cleaning script removes the brackets but keeps the text between the brackets. In this way, rather clean utterances result. The cleaned utterances are given as input to PaQu.

8. Evaluation

The assignment of parses by the Alpino parser is a fully automated process, and therefore the results will most likely contain errors. As a first small experiment to investigate the quality of the parses produced by Alpino in this domain, I used the manually tagged Van Kampen adult utterances as gold standard to compare the output of Alpino with.[28]

Alpino makes finer morpho-syntactic distinctions than the gold standard. These finer distinctions were mapped to categories distinguished in the gold standard.[29] When comparing the Alpino results with the gold standard, two types of errors were found in the manual annotations:[30]

1. Some utterances contain multiple occurrences of the relevant words but only one of them was tagged. For this case, the manual annotations were extended with the missing ones, and the extended annotations were used as gold standard.

2. Some manual annotations were simply wrong. In order to take these cases into account, a second gold standard was created in which these errors were corrected.

A striking first finding is that *heel* occurs much more often than *erg*, which occurs much more often than *zeer*. This holds for the words themselves (in the proportion 76% - 17% -7%) and even more for their use as a modifier of A,V, or P (proportion 90% - 9% - 1%). The results of the comparison with the reference annotations are provided in Table 2. It specifies the accuracy of the Alpino parser in characterizing the words *heel*, *erg* and *zeer* correctly compared to the original gold standard (Acc) and the revised gold standard (RevAcc):
The results for *heel* and *erg* are very good with over 90% accuracy compared to the revised gold standard. The results of *zeer* appear to be very bad. Further analysis reveals that most errors are made for the construction *zeer doen*, lit. *pain do*, 'to hurt', which Alpino really does not know how to analyze. The word *zeer* in this expression is correctly analyzed by Alpino as a noun, an adjective, or an adverb[31], but the grammatical functions assigned vary widely and are mostly incorrect: *direct*

[27] Of course, the final version of PaQu should natively support the CHILDES CHAT format as an input format and do the clean-up automatically. See section 11.

[28] If one logs in into the PaQu application, one actually finds the parsed corpora with the cleaned Van Kampen adult sentences, since I shared the corpora with everyone. They are called *VanKampenHeel*, *KampenErg*, and *VanKampenZeer*, resp.

[29] Alpino classifies the word *wat* as a pronoun, in accordance with traditional grammar and the conventions for part of speech tagging adhered to by Alpino (Van Eynde 2005). I mapped all occurrences where *heel* modifies a pronoun to *mod A*, which is wrong for *wat* (it is clearly a noun). This, however, does not affect the accuracy scores.

[30] Typical examples of human errors, mostly caused by sloppiness, lack of concentration, etc.

[31] Alpino distinguishes adverbs from adjectives in some cases by means of the syntactic category. The gold standard does not distinguish adverbs from adjectives by syntactic category.

word	**Acc**	**RevAcc**
heel	0.94	0.95
erg	0.88	0.91
zeer	0.21	0.21

Table 2: Accuracy of Alpino parses for adult utterances in the CHILDES Van Kampen subcorpus

object, *predicative complement*, *modifier*, and even *subject*. For a linguist, the analysis is also not obvious, but I have analyzed *zeer* in this construction in all cases as a predicative complement to the verb *doen*. Whether *zeer* is a noun or an adjective is often indeterminable, and this distinction has not been taken into account in making the comparison.

Since the bad results for *zeer* are mainly caused by one type of construction, which can be easily identified in PaQu[32], the results of PaQu are still very useful.

Though the results for *heel* and *erg* are very good, several caveats must be made. First, these utterances have been cleaned, so that actually idealized utterances are parsed. I believe this step is justified, and probably needed to get useful results, but one has to remain aware of this idealisation. Second, most utterances are relatively short, and most of the sentences are explicitly separated from each other in the CHILDES corpora: splitting a running corpus into a sequence of sentences is a non-trivial and error-prone process. Third, most grammatical relations investigated are very local (an adjective modifying an adjective or preposition is usually adjacent to its modifiee). Such local grammatical relations may be analyzed correctly even in sentences that are overall assigned completely wrong parses. Finally, the experiment here involves adult speech. These considerations make it clear that one cannot simply generalize the results achieved here to other cases. Nevertheless, the results reported here are promising.

In fact, I actually also evaluated the performance of Alpino on the children's utterances. We found similar results, though the accuracy figures are lower. The relative proportion of *heel*, *erg* and *zeer* shows a distribution similar to the adults utterances, though with an even higher proportion for *heel*: (93% - 5% - 2%) and (96% - 4% - 0%) for their use as a modifier of A,V, or P. For the accuracy figures, see Table 3:

word	**Acc**
heel	0.90
erg	0.73
zeer	0.17

Table 3: Accuracy of Alpino parses for children's utterances in the CHILDES Van Kampen subcorpus

I ignored examples where the gold standard specifies *unclear* as value (23 out 333 examples, usually incomplete or ungrammatical utterances) or where the word *heel*, *erg* or *zeer* is deleted by the cleaning process (3 out 333 examples). The reasons why these figures are lower than the figures for the adult utterances are as follows. First, several of the children's sentences are ungrammatical, e.g., *de prinses is hele groot* 'the princess is very-E big' with an inflected form of *heel* (this concerns about 9 cases which have not been marked as *unclear*). Second, the annotators have represented the pronunciation of the words by the children in the orthography (e.g. *feel* instead of *veel* 'much', *sem* for *zwem* 'swim', etc.) without explicitly marking what is intended with these strings. This

[32] Through the query `http://zardoz.service.rug.nl:8067/?db=childesadultsheelerga&word=zeer&rel=&hword=%2Bdoen&postag=&hpostag=`;login is required to access the corpus.

concerns 24 utterances, for each of which the Alpino parse is wrong. The parse is usually wrong because Alpino analyzes the unknown strings as nouns. Adding annotations for these cases might make the performance slightly higher.

9. Analysis for all Dutch CHILDES corpora

The results of an analysis of the words *heel*, *erg* and *zeer*, based on an automatic parse of all adult utterances in the Dutch CHILDES corpora are given in Table 4.[33] It specifies, for each of the three words, the counts of their occurrences in specific grammatical roles that concern us here, the counts of their occurrences in other grammatical roles (*other*), and of cases where the grammatical role could not be determined (*unclear*).[34]

Results	**mod A**	**mod N**	**Mod V**	**mod P**	**predc**	**other**	**unclear**	**Total**
heel	881	51	2	2	14	0	2	**952**
erg	347	27	109	0	187	5	0	**675**
zeer	7	1	83	0	19	21	7	**138**

Table 4: Analysis of *heel*, *erg* and *zeer* in adult utterances in Dutch CHILDES

The proportion of *heel*, *erg* and *zeer* shows a similar distribution as in the Van Kampen subcorpus, though the frequency of *erg* is much higher than in the Van Kampen Corpus: 54% - 38% - 8%. In their use as a modifier of A,V, or P the proportions are 65% - 34% - 1%.[35]

Most striking in these data is the overwhelming number of cases where *heel* modifies an adjective. This covers over 92% of the examples with *heel* found. Modification of V and P by *heel* hardly occurs, and in fact the four examples all involve wrong parses. The mod V cases actually involve adjectives (*beroemd* 'famous', and *verschillend* 'different') that happen to be identical in form to verbal participles and that are always analyzed by Alpino as verbs.[36] In the mod P examples, *heel* is a secondary adjectival predicate in one case, and it is unclear how it should be analyzed in the other example.[37]

A second observation is that there are quite some examples in which *erg* modifies an adjective or a verb.

A third observation is that there are very few examples involving *zeer* modifying an adjective. In only 6 out of the 83 examples of Mod V, *zeer* indeed modifies a verb. In one case it modifies an adjective. All other examples where it modifies V according to Alpino are in fact wrong parses involving *zeer doen* 'to hurt', discussed above). The scarcity of the examples of *zeer* as a modifier of A,V, or P can plausibly be attributed to its more formal pragmatic nature, so that it will be less used in spoken parent - young child interactions.

The table suggests that there are no examples of *erg* and *zeer* modifying prepositional phrases at all. In fact, there are a few (4 occurrences), but they involve idiomatic expressions such as *op prijs stellen*[38] (modified by *erg* once and by *zeer* twice) and *in de smaak vallen*[39] (modified by *zeer* once) in which Alpino has analyzed them as modifying the verb.

[33] The results reported here deviate slightly from what (Odijk 2015) reported. In the current table the wrong mapping of the pronoun *wat* has been corrected, and changed from *mod A* to *mod N*. This concerns 5 examples, all modified by *heel*. This small correction does not affect the overall results.

[34] For example, in incomplete or ungrammatical utterances.

[35] These figures are based on reassignments for wrong parses of *heel* and *zeer*, see below.

[36] More precisely, Alpino presents an analysis in accordance with the annotation guidelines for participles (Van Eynde 2005, 26).

[37] The sentence is *daar kon je heel in lopen.*

[38] Lit. at price set, 'appreciate'

[39] lit. in the taste fall, 'like' (with arguments reversed).

10. Conclusions

We can draw two types of conclusions from the work presented in this paper: conclusions with regard to the linguistic problem, and conclusions with regard to PaQu as a research tool.

Starting with the linguistics, any conclusions here must be very preliminary, given the small scale of the research done here. Nevertheless, the observations made in the preceding section are suggestive of further research. For example, they suggest that the overwhelmingness of the occurrence of *heel* as a modifier of an adjective in comparison to its occurrence as a modifier of a verb (881 v. 2), perhaps in combination with its early occurrence[40], might play a role in fixing the modification potential of this word to adjectives. In contrast, the occurrences of the word *erg* as a modifier of adjectives and verbs are more balanced: 347 v. 109.

The fact that there are hardly any examples for *zeer* make it difficult to draw any conclusions. In any case, the current CHILDES data give no clue how the use of *zeer* as a modifier of A,V,P is acquired, simply because there are hardly any data. This most probably means that the current CHILDES samples are insufficiently large as a sample of first language acquisition.[41]

Turning to the research tool PaQu, it can be safely concluded from this paper that PaQu is very useful for aiding researchers in better and more efficient manual verification of hypotheses than without this tool. Because of its fully automated nature, it applies blindly and automatically and is in this respect usually more consistent than humans (who err). But of course, the parses generated by PaQu (via Alpino) are fully automatically generated and will contain errors. Nevertheless, as shown in this paper, in some cases, its fully automatically generated parses and their statistics can reliably be used directly (though care is required!), and one frequently occurring error described in this paper turned out to be so systematic that the relevant examples can be easily identified using PaQu queries.

11. Future Work

It is obvious that the work reported on in this paper is just the beginning. There is a lot of work that can (and should) be done in the near future. Firstly, the same words could be investigated in other corpora that are relevant for language acquisition, in particular the Basilex corpus[42]. Secondly, similar experiments can be carried out for other tuples of (near-)synonymous words with different syntactic selection or modification properties. One example is *te* v. *overmatig*, which both mean 'too' but differ in modification potential (*te* only A, *overmatig* at least A and V). Another example concerns the copular verbs *worden* 'become' v. *raken* 'get', in which *worden* can only take NP, AP and a very limited number of PP predicates, while *raken* can take only AP and PP predicates, very similar to their English translations *become* and *get*. Of course, as usual in natural language, most of these words are ambiguous.[43] Most of these ambiguities can be resolved by the syntactic contexts, so treebanks can (and must) be used to find the relevant examples and their statistics.

PaQu can be improved in many ways. Currently it only allows plain text as input, but it should actually support input in all formats commonly used in linguistics, e.g, the CHILDES CHAT format, the FoLiA[44] format (van Gompel and Reynaert 2013) and TEI[45]. In addition, it should take in not only the actual data, but also the metadata of the corpus, its subcorpora or textual units such as

[40] (Odijk 2014, 34), using the COAVA application, also developed in CLARIN-NL, observed that the first occurrence of *heel* as a modifier of A,V, or P in the children's utterances is on day 705 (1;11). The first occurrence of the word *erg* modifying A,V, or P is on day 1048 (2;10), while the first occurrence of *zeer* modifying A,V, or P is found on day 1711 (4;8).

[41] A rough count shows that the Dutch CHILDES corpora dealt with here contain 534 k utterances and approx. 2.9 million inflected word form occurrences ('tokens').

[42] `http://tst-centrale.org/nl/producten/corpora/basilex-corpus/6-158`

[43] For example, *te* is an adjective, a preposition, and an infinitive marker; *raken* is not only a copula but also a transitive verb (with two meanings); *worden* is not only a copula but also a passive auxiliary.

[44] `http://proycon.github.io/folia/`

[45] `http://www.tei-c.org/index.xml`

utterances, paragraphs etc. And it should enable users to carry out analyses not only on the data (which is currently possible in limited ways) but on arbitrary combination of search results data and and their metadata.

Furthermore, the functionality of uploading one's own corpus should also be added to other treebank search applications, in particular the GReTEL[46] application (Augustinus et al. 2012).

And finally, it would make sense to manually verify and where needed correct (parts of) parses for CHILDES corpora, improving the reliability of these data. Most of the possible future work mentioned here is actually planned in the CLARIAH-CORE[47] project or in the Utrecht University project *AnnCor*.

Acknowledgements

This work crucially uses data and/or tools made available through the CLARIN infrastructure. The work was financed by CLARIN-NL[48], an NWO project in the Netherlands. I thank the anonymous reviewers for valuable comments on an earlier version of this paper.

References

Augustinus, Liesbeth, Vincent Vandeghinste, and Frank Van Eynde (2012), Example-based treebank querying, *in* Calzolari, Nicoletta, 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 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey.

MacWhinney, Brian (2015), Tools for analyzing talk, electronic edition, part 1: The CHAT transcription format, *Technical report*, Carnegie Mellon University, Pittsburg, PA. http://childes.psy.cmu.edu/manuals/CHAT.pdf.

Odijk, Jan (2011), User scenario search, internal CLARIN-NL document. http://www.clarin.nl/sites/default/files/User%20scenario%20Serach%20110413.docx.

Odijk, Jan (2014), CLARIN: What's in it for linguists? Uilendag Lecture, Utrecht. http://dspace.library.uu.nl/handle/1874/295277.

Odijk, Jan (2015), Linguistic research with PaQu, lecture held at CLIN 2015, Antwerp, http://www.clarin.nl/sites/default/files/Poster%20Odijk%20CLIN%202015%202015-02-02.pdf.

Tjong Kim Sang, Erik, Gosse Bouma, and Gertjan van Noord (2010), LASSY for beginners, Presentation at CLIN 2010. http://ifarm.nl/erikt/talks/clin2010.pdf.

van den Bosch, A., G.J. Busser, W. Daelemans, and S. Canisius (2007), An efficient memory-based morphosyntactic tagger and parser for Dutch, *in* Van Eynde, F., P. Dirix, I. Schuurman, and V. Vandeghinste, editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, pp. 99–114.

van der Beek, Leonoor, Gosse Bouma, and Gertjan van Noord (2002), Een brede computationele grammatica voor het Nederlands, *Nederlandse Taalkunde* **7**, pp. 353–374.

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

[46] http://nederbooms.ccl.kuleuven.be/eng/gretel

[47] http://www.clariah.nl

[48] http://www.clarin.nl

van Gompel, Maarten and Martin Reynaert (2013), FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study, *Computational Linguistics in the Netherlands Journal* **3**, pp. 63–81. http://www.clinjournal.org/sites/clinjournal.org/files/05-vanGompel-Reynaert-CLIN2013.pdf.