A Use Case for Linguistic Research on Dutch with CLARIN

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

Abstract

In this paper I describe a particular Dutch linguistic problem and I show that it can be addressed in a better, more efficient, and more user-friendly manner than ever before, thanks to CLARIN. Most of the data that are used in the investigation could only be used by technical experts a few years ago but are now available to all linguists through a variety of easily accessible web applications developed in CLARIN with interfaces dedicated to their intended users. However, it also shows that still a lot of further extensions and improvements can and must be made. Fortunately, most of these are being implemented in currently running projects.

1 Introduction

In this paper I describe a particular Dutch linguistic problem and I show that it can be addressed in a better, more efficient, and more user-friendly manner than ever before, thanks to CLARIN. Most of the data that are used in the investigation could only be used by technical experts a few years ago but are now available to all linguists through a variety of easily accessible web applications developed in CLARIN with interfaces dedicated to their intended users, linguists.[1]

The relevant problem was first defined in unpublished work (Odijk, 2011). This report also specified what kinds of search actions would be needed to address this problem. At the time, almost none of these search actions were possible, or only with great difficulty, and they required expert knowledge on the relevant resources and programming or scripting skills. In 2014, (Odijk, 2014a) showed in a lecture that many of the desired search actions had become possible, in a simple manner, and through applications with interfaces dedicated to the targeted users, linguists. At the same time, it was observed that not everything was possible yet in an easy way, and new requests arose by using the relevant applications. Since neither (Odijk, 2011) nor (Odijk, 2014a) was published, I report on their findings in this paper, and I will show new functionality created to accommodate the newly arisen needs. This paper thus serves as an example of a report on a *research pilot*: a project to use functionality offered by the infrastructure with the twin goals of furthering the research but also of identifying novel functionality that the infrastructure should offer to be able to further the research.

I introduce the basic facts to be investigated in section 2, make an assessment of these facts in section 3, and list a few of the many research questions that these facts raise in section 4. I then show that a variety of web applications developed in CLARIN for searching in linguistic resources (lexicons and corpora), for enriching corpora and for analysing search results make research into this problem possible that is based on more data, which are found faster and easier than was possible ever before. The web applications considered are OpenSONAR (section 5.1), the LASSY Word Relations Search engine (section 5.2), GRETEL (section 5.3), CORNETTO (section 5.4), COAVA (section 5.5), and PaQU (section 5.6). All applications mentioned are available in the CLARIN infrastructure and can be accessed via the CLARIN-NL portal[2]. I summarize the conclusions in section 6.

[1] CLARIN as a whole of course targets all humanities researchers, but the applications discussed in this paper target linguists.

[2] http://portal.clarin.nl/

This paper shows that great progress has been made since 2011 in the number of applications offered in the CLARIN infrastructure, and it shows a significant increase in the functionality that they offer, but it also identifies functionality that was desired from the start as well as novel desired functionality that have not been implemented yet. Section 7 describes future work that can and must be done to address the research questions.

2 Basic facts

In this section I introduce the basic facts related to the problem that I want to investigate. It is a specific case of the problem of the acquisition of lexical properties by first language (L1) learners.

The three Dutch 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*. The same holds in (1c) for the verbal predicate *verbaast*.[3] 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
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. Some of these will be discussed below. I refer to (Odijk, 2011), (Odijk, 2014a) and (Odijk, 2015a) for further details.

3 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.[4] 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*. (Odijk, 2015a, section 4) describes these differences. 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, (Odijk, 2015a, section 4) shows that this is not the case for any of these differences.

[3] or maybe the whole VP *verbaast mij*.
[4] See (Odijk, 2011) for more examples supporting this conclusion.

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.

4 Research questions

The simple facts described in the preceding sections are interesting for a number of reasons. First, they constitute a kind of minimal pair in first language acquisition: though *heel* on the one hand and *zeer*, *erg* on the other are (near-)synonyms, their syntactic modification potential differs. They also illustrate acquisition of a negative property: L1 learners must learn that *heel* canNOT modify verbal or prepositional predicates. These facts therefore raise many research question related to language acquisition. Examples of these research questions are:

(3) 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 (i.e., absence of evidence interpreted as evidence for absence)?

Obviously, this paper cannot address all these questions. The main purpose of this paper is to show that, by using CLARIN, research questions such as the ones in (3) can be addressed in a better and more efficient manner than without CLARIN. In this paper, I will focus on research question (3c)

In order to address these research questions, data are needed that can provide evidence on these questions. Fortunately, many such data exist. We will mention several relevant sets in the coming sections. However, though most of these data existed before CLARIN, they were hardly usable for supporting linguistic research at the time.

5 Search and Analysis with CLARIN web applications

I described the problem of section 2 in (Odijk, 2011), as an example user scenario for search applications to be developed in CLARIN. At the time, many of the search actions I would like to be able to carry out were not possible yet, or could only be carried out with great difficulty and only with expert knowledge of the relevant data sets and query options. Some queries suggested there involve search in metadata, an area where much progress has been made since then, though some of the specific queries suggested are still not possible (and there are many other problems with searching for data via metadata through the Virtual Language Observatory[5], as described in (Odijk, 2014b)). We will not discuss this here anymore. Other suggested queries involve search in the data themselves. I list most of them here, together with an indication where they will be dealt with in this paper:

- search for synonyms, hyponyms, and co-hyponyms for the words *heel*, *erg* and *zeer* (discussed in section 5.4)
- search for bi-grams in corpora with linguistic annotations on tokens (discussed in section 5.1)
- search in the Dutch CHILDES corpora, in the children's speech, and the speech by adults addressing children (discussed in section 5.5)
- search in treebanks (discussed in sections 5.2 and 5.3)
- search in CHILDES corpora enriched with syntactic structures / PoS-tags (discussed in section 5.6)

[5] `https://vlo.clarin.eu/?0`

In March 2014, I investigated what was possible at the time, and reported on that in a lecture (Odijk, 2014a). Some crucial functionality which was still lacking was identified, which led to plans for the creation of two new applications, *PaQu* (see section 5.6) on which (Odijk, 2015a) reported extensively, and *AutoSearch* (briefly discussed in section 5.6). The easiest way to get a first overview of what kind of applications developed in CLARIN-NL can be used for humanities research is via the CLARIN-NL portal[6], which allows faceted search by *research domain*, *tool task*, *language* and other facets. For a more detailed assessment of the suitability of a certain application for a specific research question, the application has to be studied in more detail through its documentation or via a tutorial (see the CLARIN-NL portal's *Educational Packages Section* for available educational material.)

Several suggested queries can be now carried out, but many are not yet possible. We will take up this issue in section 7.

5.1 OpenSONAR

OpenSONAR[7] is a web application that enables search in and analysis of the large scale Dutch reference corpus SONAR[8] and SONAR New Media[9] (Oostdijk et al., 2013). In part because of the size of the corpus (500 million tokens[10]), accessing the information contained in the data set has proven to be difficult. OpenSONaR facilitates the use of the SoNaR corpus by providing a user-friendly online web interface tuned to the intended users, linguists. No software or data need to be downloaded, no programs installed, and no programming knowledge is required.

SONAR is a reference corpus of contemporary written Dutch for use in different types of linguistic (incl. lexicographic) and HLT research and the development of applications. It was created in the STEVIN (Spyns and Odijk, 2013) funded SONAR project (2008-2011) that built on the results obtained in the earlier STEVIN projects D-Coi and Corea.

SONAR contains over 500 million tokens of full texts from a wide variety of text types from conventional media. SONAR New Media contains texts from the social media (Twitter, Chat, SMS) with about 35 million tokens. These corpora have been tokenized, tagged for part of speech and lemmatized, and Named Entities have been labelled. All annotations were produced automatically, no manual verification took place.

OpenSONAR is an online application for exploration of and searching in the SoNaR corpus. In the *Exploration* interface one can look into the corpus distributions, request statistics from sub-corpora, retrieve n-grams from sub-corpora and search for specific documents using the SoNaR document ID. In the *Search* interface one can use any of four different search strategies: simple, extended, advanced or expert .

OpenSONAR makes it easy to search for two adjacent tokens (bigrams) via their properties *lemma*, *word*, and *part-of-speech* (*pos*). For example, one can search for a token with lemma="heel" immediately followed by a token with pos="preposition", or the same with lemma="zeer" instead of "heel", or for a token with lemma="heel" immediately followed by a token with pos not equal to adjective.

Adjacency of tokens does of course not imply a grammatical relation of modification. Therefore the search results will contain many false hits. Nevertheless the search results are useful, in particular because the search results can be sorted and grouped in various ways, which reduces the effort of separating correct from false hits.

Analysis of the search results yield several new results. Firstly, it turns out that *heel* **can** modify certain PPs, in particular certain adverbial PPs, such as

(4) a. heel in het begin
very in the beginning

[6] http://portal.clarin.nl/

[7] http://portal.clarin.nl/node/4195

[8] https://vlo.clarin.eu/record?q=SONAR&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1498_64_format_61_cmdi

[9] https://vlo.clarin.eu/record?q=SONAR&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1502_64_format_61_cmdi

[10] I use the term *token* in this paper as a term for *occurrence of an inflected word form.*

in the very beginning

b. heel af en toe
very off and to
very infrequently

c. heel in het bijzonder
very in the particular
very specially

d. heel op het laatst
very on the last
at the very last moment

e. heel in de verte
very in the distance
very far away

f. heel uit de verte
very from the distance
from very far away

g. heel in het algemeen
very in the general
very generally

These examples do not undermine our earlier claims on the data, but do add a new set of data that clearly should be incorporated in the analysis.

Secondly, *heel* does indeed also occur with predicative PPs in SONAR as in (5):

(5) a. heel buiten zijn verwachting
very outside his expectation
completely unexpectedly

b. heel in de mode
very in the fashion
completely fashionable

c. heel in de vakantiestemming
very in the holiday-mood
completely in the mood for a holiday

d. heel in het zwart
very in the black
completely without paying taxes

e. heel in orde
very in order
completely OK

I find all examples of (5) ill-formed. The mere occurrence of such examples in a corpus need not be significant, since corpora contain examples of actual language use, which may contain errors. However, their number is sufficiently large to suspect that we are dealing here with a genuine instance of variation in the Dutch language. Though I glossed the word *heel* here as *very*, I think that people who use such expressions intend *heel* here in the sense of *geheel* or *helemaal* ('completely'), and the translations I provided in (5) reflect this. Obviously, one would like to investigate further properties of these utterances (e.g., genres that they occur in, origin of the utterer (Netherlands or Flanders), his/her gender and age etc.), but that is not so easy with the current version of OpenSONAR: The search output contains many false hits. Though one can cross-classify *all* search results with metadata information, one cannot mark a subset of search results for such a cross classification with metadata. An extension of OpenSONAR is required for this (see section 7).

5.2 LASSY Word Relations Search Engine

As mentioned above, adjacency of two words does not imply that these two words entertain a grammatical dependency. What one would actually want is a database in which grammatical dependencies between words are represented and are searchable. This information is available in treebanks, but the databases that contain this information are much smaller than SONAR. The LASSY Word relations Search Engine (LWRS)[11] (Tjong Kim Sang et al., 2010) enables one to search for such grammatical dependencies in certain treebanks. Actually, LWRS already existed when I described the linguistic problem for the first time. It was originally not developed in the CLARIN-NL project, but clearly inspired by the desire in CLARIN to provide web applications for search in corpora with interfaces that are tuned to linguists as users.

LWRS has a dedicated interface that enables a user to specify a query that searches for utterances containing two words entertaining a grammatical dependency by providing the properties of these two words (lemma, word form, part of speech) and the label of their grammatical dependency (e.g. subject, object, etc.).[12] This makes it easy to search for utterances that e.g. contain a word with lemma *heel* that is a modifier of a word with pos *verb*, and many similar examples.

Such queries carried out on the 1 million token manually verified written language treebank LASSY-Small Corpus[13] (van Noord et al., 2013) yield the following results:

- LASSY-SMALL contains examples where *heel* appears to modify a verb, but in all cases these are adjectives that happen to be identical in form to the participles of verbs. In such cases, LASSY, by convention, always analyzes these as verbs.

- LASSY-SMALL incorrectly analyzes *heel* in *heel open staan for* lit. very open stand for, 'be very receptive for' as modifying the verb *staan*, while in fact it modifies the adjective *open*.

- LASSY-SMALL contains examples where *erg* or *zeer* modifies verbs. In most cases, this also involves adjectives that happen to be identical to participles of verbs, but there are also several cases of modification of a real verb.

In short, these findings confirm our initial assumptions of the facts, which are now backed by a large amount of empirical material.

5.3 GRETEL

GrETEL is web application that enables a user to provide an example sentence of a construction that he/she is interested in and to specify which aspects of this example sentence are crucial for identifying the construction. The system then automatically generates a query and applies it to a treebank (LASSY-SMALL or the Spoken Dutch Corpus treebank, each manually verified and containing 1 million tokens).[14]. The query is generated by parsing the example sentence with the same parser that was used in the creation of the treebank (Alpino[15] (van der Beek et al., 2002)), thus increasing the chances of providing a query that finds instances of the construction searched for. The GrETEL application has been described in detail elsewhere (Augustinus et al., 2012; Vandeghinste and Augustinus, 2014).

Applying it to the Spoken Dutch Corpus (Oostdijk et al., 2002) yields the following results:[16]

- The word *heel* occurs as a modifier of a verb in 61 cases. However,

[11] `http://portal.clarin.nl/node/1966`

[12] Not specifying any properties matches with every word or relation, so this functions basically as a variable in the query.

[13] `https://vlo.clarin.eu/record?q=LASSY-SMall&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1493_64_format_61_cmdi`

[14] And since recently, also the automatically parsed SONAR-500 corpus.

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

[16] The full GrETEL functionality is not necessary for the problem at hand, though it can be used for it. In fact, the analysis described here has been carried out with PaQu (see section 5.6), since its options for analyzing the search results are more extensive than GrETEL's. Extension of GrETEL's analysis options is planned for the future. See section 7.

  - in 53 of these, the word is actually an adjective that happens to be identical to the participle of a verb (as above in LASSY-Small);
  - in 3 cases *heel* actually modifies a substantivised infinitive (and, as as modifier of a noun, has the meaning 'whole');
  - in 2 cases I find the sentence ill-formed. Maybe *heel* is intended here as 'completely'. Both utterances are of Flemish origin;
  - in 3 cases the analysis in the treebank is incorrect;

- The word *heel* occurs as a modifier of a preposition in 6 cases:
  - in 4 cases these are adverbial PPs that we also encountered with OpenSONAR (see section 5.1, the examples in (5);
  - in one case I find the sentence ill-formed. Maybe *heel* is intended here as 'completely'. The utterance is again of Flemish origin;
  - in one case *heel* modifies the expression *voor de hand liggen* lit. in-front-of the hand lie, 'be obvious'. I find the example marginal, except when the verb in the expression is a present participle. In that case, however, we are arguably dealing with an adjectival expression.[17]

- The word *heel* occurs as a modifier of an *MWU* (multi-word unit). These MWUs have no other part of speech code, but further analysis shows that they involve
  - adjectives in 3 cases[18];
  - nouns in 4 cases (e.g. *heel Den Haag*, lit. whole the Hague) and *heel* means 'whole' in these cases;
  - adverbial prepositional phrases in 2 cases (*heel af en toe*, lit. very off and to, 'very infrequently')
  - incorrect analyses in 3 cases

  In summary, these facts are consistent with our findings on the basis of OpenSONAR and with our initial assumptions on the data, and they suggest that the use of *heel* as a modifier of predicative PPs might be possible for certain Flemish speakers.

5.4 CORNETTO

(Odijk, 2011) suggested that analysing the modification potential of (near-)synonyms, co-hyponyms, and hyponyms of the words *heel*, *erg* and *zeer* may contribute to an understanding of the problem at hand. At the time, searching for synonyms or near-synonyms, let alone for words with other semantic relations for a given word, was very difficult. Obviously, one would want to use the Cornetto database for this purpose.

The Cornetto database is a lexical resource for the Dutch language which combines two resources with different semantic organisations: the Dutch Wordnet with its synset organisation and the Dutch Reference Lexicon which includes definitions, usage constraints, selectional restrictions, syntactic behaviours, illustrative contexts, etc. The Cornetto database contains over 92K lemmas and almost 120K word meanings.

At the time, an interface to Cornetto existed, but it often did not work, required an old version of the Firefox browser[19], and the interface itself was not well-designed. Searching for semantically related

[17] For example, it can be used predicatively and be modified by *te* 'too'

(1) Dat is te voor de hand liggend
That is too in-front-of the hand lying
That is too obvious

which is not possible for verbal present participles.

[18] In *heel ver weg*, lit. very far away, *ver weg* is analyzed as a MWU, though clearly here *heel* modifies the adjective *ver*, and together they modify the word *weg*.

[19] Arguably, this is a defect of Firefox. Producing upgrades that are not backwards compatible should be banned!

words has become easy with CLARIN, since a web application with a dedicated interface to the Cornetto database has been created.

The Cornetto web application[20] offers 3 different interfaces: Simple search for lexical entries[21], Advanced search for lexical entries[22], and Search for synsets[23].

Searching for (near-)synonyms of *zeer* in the relevant sense (Cornetto sense identifier *zeer-adv-3*) yields the following set of sense identifiers from Cornetto:[24]

(6) allemachtig-adv-2, beestachtig-adv-2, bijzonder-a-4, bliksems-adv-2, bloedig-adv-2, bovenmate-adv-1, buitengewoon-adv-2, buitenmate-adv-1, buitensporig-adv-2, crimineel-a-4, deerlijk-adv-2, deksels-adv-2, donders-adv-2, drommels-adv-2, eindeloos-a-3, enorm-adv-2, erbarmelijk-adv-2, fantastisch-adv-6, formidabel-adv-2, geweldig-adv-4, goddeloos-adv-2, , godsjammerlijk-adv-2, grenzeloos-adv-2, grotelijks-adv-1, heel-adv-5, ijselijk-adv-2, ijzig-a-4, , intens-adv-2, krankzinnig-adv-3, machtig-adv-4, mirakels-adv-1, monsterachtig-adv-2, moorddadig-adv-4, oneindig-adv-2, onnoemelijk-adv-2, ontiegelijk-adv-2, ontstellend-adv-2, ontzaglijk-adv-2, ontzettend-adv-3, onuitsprekelijk-adv-2, onvoorstelbaar-adv-2, onwezenlijk-adv-2, , onwijs-adv-4, overweldigend-adv-2, peilloos-adv-2, reusachtig-adv-3, reuze-adv-2, schrikkelijk-adv-2, sterk-adv-7, uiterst-adv-4, verdomd-adv-2, verdraaid-a-4, verduiveld-adv-2, , verduveld-adv-2, verrekt-adv-3, verrot-adv-3, verschrikkelijk-adv-3, vervloekt-adv-2, vreselijk-adv-5, waanzinnig-adv-2, zeer-adv-3, zeldzaam-adv-2, zwaar-adv-10

The word *heel*, in one of its senses (with Cornetto sense identifier *heel-adv-5*) is included here.

Similarly, the near-synonyms of *erg*, in the relevant sense (with Cornetto sense identifier *erg-a-2*) are listed in (7):

(7) erg-a-2, ernstig-a-2, fel-a-1, hard-a-4, heftig-a-1, hevig-a-1, krachtig-a-3, sterk-a-4, stevig-a-2, straf-a-2, vet-a-5, vurig-a-1, zwaar-a-3

And the hyponyms of these senses can be retrieved easily as well. Cornetto thus offers, in a very simple way, a list of word senses (and therefore words) that are semantically related to the word sense queried.

Now one would like to use the corpus search interfaces described above to investigate the modification potential of the words associated with these meanings. This is possible, but currently requires making a separate query for each of the words associated with the meanings listed above (some 70 words). One can also write a single query with each of the relevant words as an alternative, but the analysis options of the current corpus search and analysis applications do not enable e.g. a grouping by the modifier lemma and the modifiee part of speech. For example OpenSONAR's analysis options enable one to group the results by the part of speech of the immediately adjacent word but do not allow sorting the results by the lemmas searched for at the same time. An alternative approach, suggested by (Odijk, 2011), is parameterized search, but this has not yet been implemented in any of the search applications (see section 7).

The analysis of the modification potential of these words is therefore work for the future. It is already clear that many of the synonyms are untypical for children and are probably acquired rather late. It is therefore interesting to investigate whether there is a relation between the timing of acquisition of these words and their modification potential.

5.5 COAVA

Since the problem we are interested in concerns (first) language acquisition, it is obvious that data that directly concern language acquisition must be taken into account. The most important data set for language acquisition is the CHILDES data set.

[20] http://portal.clarin.nl/node/1944

[21] http://cornetto.clarin.inl.nl/simple_search.xql

[22] http://cornetto.clarin.inl.nl/advanced_search.xql

[23] http://cornetto.clarin.inl.nl/wordnet.xql

[24] It is pretty difficult and often quite arbitrary to add translations to these words, and they are not needed for understanding the current paper, so I left them out.

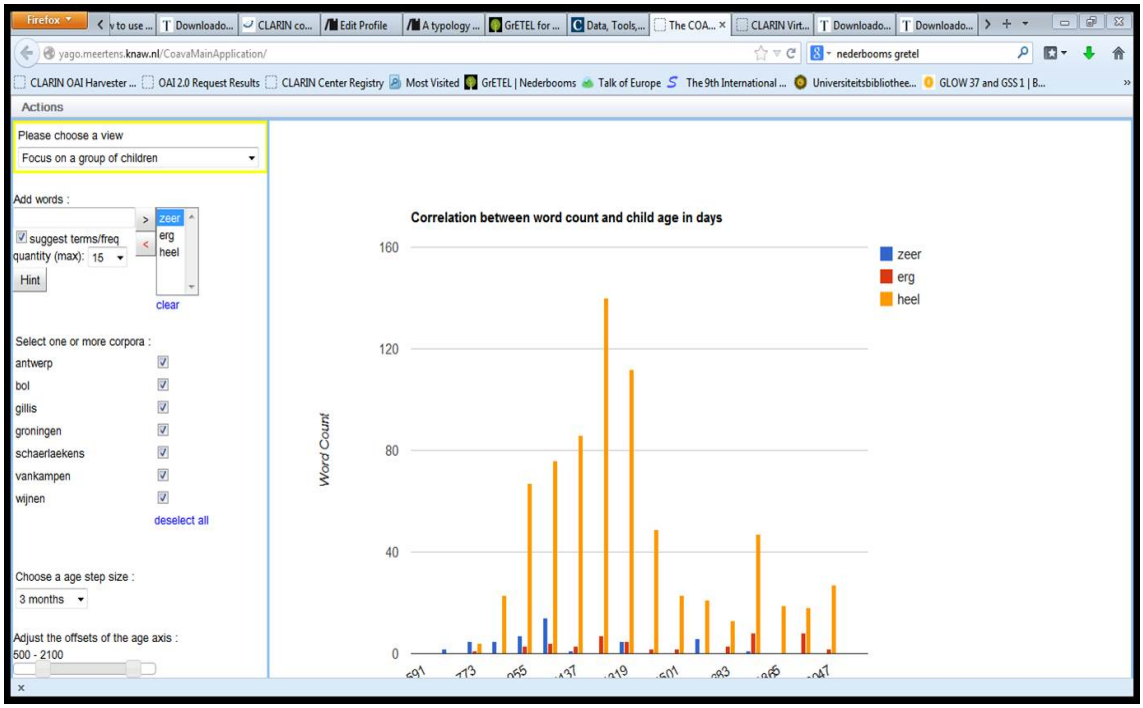

Figure 1: Frequency of the word *heel*, *erg* and *zeer* in the children's speech in the Dutch CHILDES corpora. The X-axis specifies time intervals of three months, the Y-axis indicates the frequency of the word. Each word has a separate color (blue=*zeer*, orange=*heel*, red=*erg*)

The Dutch CHILDES corpora[25] are accessible via the CLARIN Virtual Language Observatory (VLO)[26] or directly via Talkbank[27] and contain relevant material to investigate the research questions formulated in section 4. 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.

I investigated the occurrence of the words *heel*, *erg* and *zeer* in the CHILDES data through the COAVA web application[28]. The COAVA[29] web application provides combined access to two sets of databases: one with historical dialect data (the databases WBD[30] and WLD[31] with lexical data of the Brabantish and Limburgian dialect between 1880-1980) and one with first language acquisition data.

Though COAVA offers many facilities for research into the relation between language acquisition and lexical variation, my main interest is in the occurrence, and especially the first occurrence of the words *heel*, *erg* and *zeer* in the children's utterances. Figure 1 shows this.

From this figure, we can conclude that the word *zeer* occurs first, followed by *heel*, and *erg*. However, each of the words *heel*, *erg* and *zeer* is ambiguous. COAVA does not take this into account, so we do not know whether the first occurrences observed concern the relevant sense ('very') of these words. In (Odijk, 2014a) I therefore made a manual analysis, which yields different results, as shown in Table 1.[32]

From this table, one can conclude that the first occurrence of *heel* in the sense 'very' is used very early by children (before their second birthday); the first occurrence of *erg* appears only about a year later,

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

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

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

[28] `http://portal.clarin.nl/node/1928`

[29] Acronym for *Cognition, Acquisition and Variation Tool*

[30] `https://vlo.clarin.eu/search?3&fq=collection:Dictionary+of+the+Brabantic+dialects`

[31] `https://vlo.clarin.eu/search?2&fq=collection:Dictionary+of+the+Limburgian+dialects`

[32] The table specifies the age of the child in days, followed by the CHILDES notation for children's ages in the format *(year;month)*.

First Occurrence	heel	erg	zeer
Day (Year;Month)	705 (1;11)	1048 (2;10)	1711 (4;8)

Table 1: First Occurrence of *heel*, *erg* and *zeer* in the relevant sense ('very') in the Dutch CHILDES Children's speech

and *zeer* occurs only very late (far in the fourth year). The latter may be related to the fact that *zeer* is considered rather formal by many people, and also occurs rather infrequently in adult child interactions in CHILDES. Note that the very early occurrence of *zeer* in Figure 1 involves a different sense of this word, viz. as *pain* or *painful*.

Clearly, it is desirable to have the manual analysis carried out here supported or even completely replaced by an automatic procedure. The next section describes a first step towards this goal.

5.6 PaQU

As we saw in the preceding section, a serious problem for the investigation is that the words being investigated are, as any decent word in natural language, highly ambiguous. Table 2 describes the ambiguity. For example, the word *heel* is 6-fold ambiguous. This ambiguity is partly solved by taking into account morpho-syntactic and syntactic factors. For *heel* as a finite verb (Vf) the ambiguity reduces to 2, which cannot be further resolved by morpho-syntax 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'.[33]

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 2: Ambiguity of the words *heel*, *erg* and *zeer*

[33] 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., common gender) and *neuter*. *Vf* stands for *finite verb form*.

The Dutch CHILDES corpora do not contain any information about the meanings of its word occurrences. Fortunately, as is clear from Table 2, most of the ambiguities can be resolved by taking into account morpho-syntactic and syntactic properties of the word occurrences. However, as observed above, the Dutch CHILDES corpora do NOT have (reliable) morpho-syntactic information (part of speech tags) or syntactic information for the utterances either.

One would want to be able to automatically parse the CHILDES corpora, and to upload the resulting treebank in a search and analysis application. PaQu was developed for this purpose.[34]

The web application PaQu[35] was 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 LASSY Word Relations Search application[36] originally developed by (Tjong Kim Sang et al., 2010) and discussed in section 5.2), as well as via PaQu's XPATH interface.

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 (see Figure 2). The relevant queries are not easily expressed in XPATH[37], 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 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).[38] See Figure 3. 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[39] format. It currently does not allow a CHILDES corpus (in CHAT format (MacWhinney, 2015)) directly as input. This clearly requires an extension of PaQu (see section 7). For the experiments described below I wrote an ad-hoc script to select and clean utterances from CHILDES corpora (see (Odijk, 2015a) for details).

PaQu offers full parses of sentences in a corpus, but these parses have been generated in a fully automatic manner, so they will contain errors. It is therefore required to evaluate the quality of the automatically generated parses. (Odijk, 2015a) describes the results of such an evaluation for the words *heel*, *erg* and *zeer* dealt with here in the CHILDES Van Kampen subcorpus.[40] The results are summarized in Table 3, both for the adult speech (column *Adults*) and for the children's speech (column *Children*)

The results for the adults' speech and the children's speech shows a similar distribution, though the results for the children's speech are lower. For the adult speech, the results for *heel* and *erg* are very good with over 90% accuracy compared to the 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[41], but the grammatical functions assigned vary widely and

[34] Analogously, the *AutoSearch* application was developed to support search in corpora with annotations on tokens. AutoSearch 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 (van den Bosch et al., 2007) part of speech tagger in TTNWW) can be uploaded and searched via a Corpus of Contemporary Dutch -like search interface. This application will not be discussed in this paper any further.

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

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

[37] 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.

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

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

[40] 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. The children's utterances in Van Kampen are in the corpus *VanKampen-child-heelergzeer*.

[41] 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.

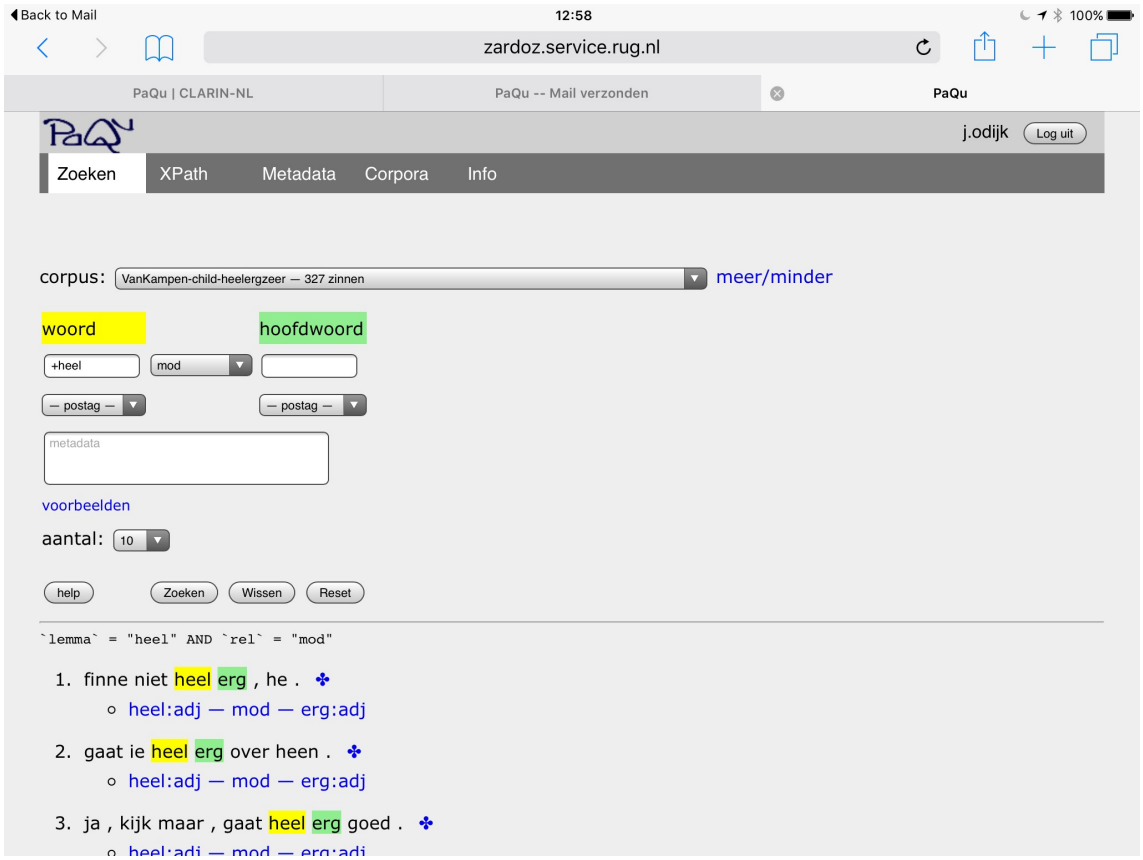

Figure 2: PaQu web interface with a query for occurrences of the lemma *heel* as modifier

word	Adults	Children
heel	0.95	0.90
erg	0.91	0.73
zeer	0.21	0.17

Table 3: Accuracy of Alpino parses for the words *heel*, *erg* and *zeer* in the CHILDES Van Kampen subcorpus

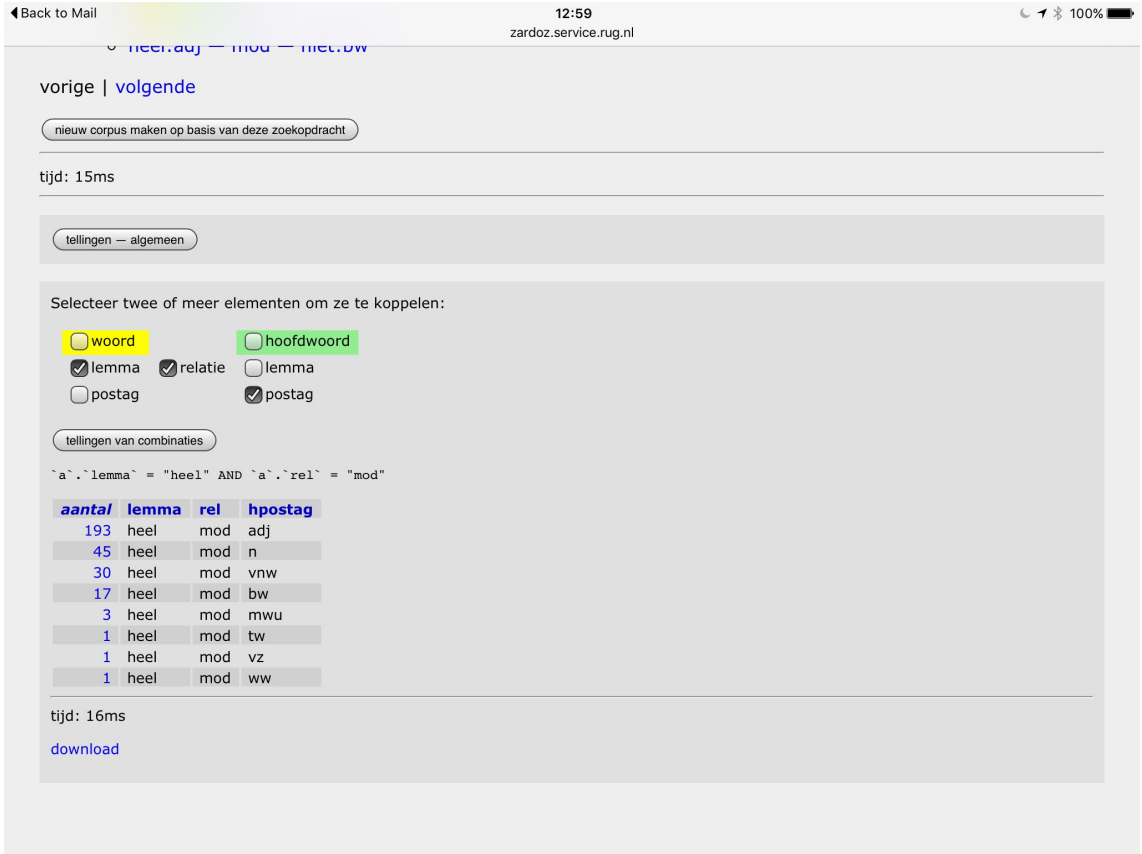

Figure 3: PaQu analysis: count of occurrences of the lemma *heel* as modifier by part of speech of the modifiee.

are mostly incorrect: *direct 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[42], the results of PaQu are still very useful.

Though (Odijk, 2015a) correctly warns against generalizing these results to other cases, they are nevertheless promising: high accuracy in some cases, and the low accuracy examples are easily identifiable.

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.[43] 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*).[44]

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

(Odijk, 2015a) analyzes these findings in some detail, and the results can be summarized as follows:

- *Heel* is most frequent (almost 54%)
- *Heel* as mod A is overwhelming: ($>$ 93%)
- *Heel* as mod V, mod P are analyzed incorrectly
- For *erg*, the distribution between Mod A and mod V is more balanced than for *heel*
- Evidence for *zeer* is mostly lacking. The examples of *zeer* as Mod V are mostly wrong analyses
- Evidence for Mod P is mostly lacking, though there is some evidence for *erg* en *zeer* (4 occurrences)

This example clearly shows the advantages of using PaQu for manual verification of hypotheses, and shows that, if some care is exercised, it can also be used for automatic verification of hypotheses. However, PaQu, in its current state, is not yet able to derive Table 1 or a variant of Figure 1 for the words *heel*, *erg* and *zeer* in the relevant sense. That requires an analysis of the search results in terms of a mix of linguistic annotations and metadata pertaining to the whole utterance or the whole session. See section 7.

6 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 CLARIN as a research infrastructure.

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 overwhelming amount of occurrences 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 (see section 5.5), might play a role in fixing the modification

[42]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.

[43]The results reported here deviate slightly from what (Odijk, 2015b) 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.

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

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 Dutch CHILDES databases are insufficiently large as a sample of first language acquisition.[45]

Concerning CLARIN, (Odijk, 2011) defined a linguistic problem and specified what kinds of search actions would be needed to address this problem. At the time, almost none of these search actions were possible, or only with great difficulty, and they required expert knowledge on the relevant databases and programming skills. In 2014, (Odijk, 2014a) showed that many of the desired search actions had become possible, in a simple manner, and through applications with interfaces dedicated to the targeted users, linguists. At the same time, it was observed that not everything was possible yet in an easy way, and new requests arose by using the relevant applications. Since neither (Odijk, 2011) nor (Odijk, 2014a) was published, I report on their findings in this paper, and I showed new functionality created to accommodate the newly arisen need. This paper thus serves as an example of a report on a *research pilot*: a project to use functionality offered by the infrastructure with the twin goals of furthering the research but also of identifying novel functionality that the infrastructure should offer to be able to further the research.

This paper shows great progress in the number of applications offered in the CLARIN infrastructure, and a significant increase in the functionality that they offer, but I have also identified functionality that was desired from the start as well as novel desired functionality that have not been implemented yet.

7 Future Work

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[46]. 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.[47] 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.

It surely also makes sense to manually verify and where needed correct (parts of) parses for CHILDES corpora, improving the reliability of the annotations on these data.

I have identified many instances of desired functionality that is not available yet. (Odijk, 2011) suggested parameterized search, but this has not yet been implemented. The functionality of uploading one's own corpus should also be added to other treebank search applications, in particular the GReTEL[48] application (Augustinus et al., 2012). All search engines that allow uploading one's own corpus must be extended to support input in all formats commonly used in linguistics. For example, PaQu only allows plain text as input, but it should actually support, e.g, the CHILDES CHAT format, the FoLiA[49] format (van Gompel and Reynaert, 2013) and TEI[50]. 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 utterances, paragraphs etc.

Search applications should offer extensive options for analyzing the search results. Such analysis options are available in PaQU and OpenSONAR, but hardly in GrETEL, and the PaQU and OpenSONAR

[45] 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').

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

[47] 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.

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

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

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

analysis options must be extended as well. In particular, the search applications should enable users to carry out analyses not only on the data but on arbitrary combinations of search result data and and their metadata.

It is also essential that the search results can be further annotated by users, or at least categorized. This is important since most search actions in practice do not yield exactly the set the researcher is interested in (there are problems of recall and of precision). With a categorisation option, one can use a broader query and then categorize the results (e.g. to exclude some).[51] And these newly added categories should participate as first class citizens in the analysis options offered.

Fortunately, most of the possible future work mentioned here is actually planned in the CLARIAH-CORE[52] project or in the Utrecht University project *AnnCor*, and part of it is already being carried out.[53] With these projects, we hope to be able to run queries such as the ones already suggested in (Odijk, 2011) but currently not possible yet (with *heel*, *erg* and *zeer* only in the relevant sense 'very'):

- For each child, give list of pairs (session, age) of the child
- For each child, give me #sessions by period, where period is e.g. every month, week, half year, year
- For each child give me the list of new words uttered by period
- For child and each session, give #occurrences of *zeer*, *heel*, *erg*;
- Idem, by period
- Give me utterances containing occurrences of *zeer*, *erg*, *heel* uttered by the child before any adult used any of these words
- Give me #occurrences of *heel* uttered by the parent before the child utters it (idem for *zeer*, *erg*, etc.)

and many others that might be needed to address the research questions of section 4.

Acknowledgements

This work crucially uses data and/or tools made available through the CLARIN infrastructure. The work was financed by CLARIN-NL[54], an NWO project in the Netherlands.

References

[Augustinus et al.2012] 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 2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[MacWhinney2015] Brian MacWhinney. 2015. Tools for analyzing talk, electronic edition, part 1: The CHAT transcription format. Technical report, Carnegie Mellon University, Pittsburg, PA, April27. `http://childes.psy.cmu.edu/manuals/CHAT.pdf`.

[Odijk2011] Jan Odijk. 2011. User scenario search. internal CLARIN-NL document, `http://www.clarin.nl/node/166`, April 13.

[Odijk2014a] Jan Odijk. 2014a. CLARIN: What's in it for linguists?, March 27. Uilendag Lecture, Utrecht, `http://dspace.library.uu.nl/handle/1874/295277`.

[51] The Lancaster web access to the British National Corpus offers such categorisation options.

[52] `http://www.clariah.nl`

[53] For example, in AnnCor manual verification and correction of Alpino parses for CHILDES utterances is worked on, and since a few months, PaQu enables analysis of search results in combination with metadata, at least for the Spoken Dutch Corpus. And it already supports more input formats than just plain text, among them FOLIA and TEI.

[54] `http://www.clarin.nl`

[Odijk2014b] Jan Odijk. 2014b. Discovering resources in CLARIN: Problems and suggestions for solutions. unpublished article, Utrecht University, http://dspace.library.uu.nl/handle/1874/303788, August.

[Odijk2015a] Jan Odijk. 2015a. Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14, December.

[Odijk2015b] Jan Odijk. 2015b. 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, February 6.

[Oostdijk et al.2002] N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In M. González Rodriguez and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.

[Oostdijk et al.2013] N. Oostdijk, M. Reynaert, V. Hoste, and I. Schuurman. 2013. The construction of a 500 million word reference corpus of contemporary written Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, pages 219–247. Springer, Berlin. http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1.

[Spyns and Odijk2013] P. Spyns and Jan Odijk. 2013. *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*. Springer. http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1.

[Tjong Kim Sang et al.2010] Erik Tjong Kim Sang, Gosse Bouma, and Gertjan van Noord. 2010. LASSY for beginners. Presentation at CLIN 2010, Utrecht, February 5.

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

[van der Beek et al.2002] Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7:353–374.

[van Gompel and Reynaert2013] Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12/2013.

[van Noord et al.2013] 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. http://dx.doi.org/10.1007/978-3-642-30910-6_9.

[Vandeghinste and Augustinus2014] Vincent Vandeghinste and Liesbeth Augustinus. 2014. Making large treebanks searchable. The SoNaR case. In Marc Kupietz, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt, and Jani Takhsha, editors, *Proceedings of the LREC2014 2nd workshop on Challenges in the management of large corpora (CMLC-2)*, pages 15–20. ELRA, Reykjavik. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-CMLC2%20Proceedings-rev2.pdf.