Jan Odijk

CLARIN's Support for Research into the Acquisition of Lexical Properties

Abstract: Odijk (2011) sketched a research question on the acquisition of lexical properties of words, and illustrated it with some concrete examples, in particular with respect to the lexical properties of the Dutch synonyms *heel*, *erg*, and *zeer* (all meaning 'very'). This work also indicated what the CLARIN infrastructure should offer to make it possible to address this research question. In this contribution I sketch to what extent the CLARIN infrastructure has achieved these requirements and desiderata. The resulting picture is mixed: (1) some have been implemented; (2) some have not been implemented and are still highly desirable; (3) some have not been implemented but turned out to be not so urgent; (4) new requirements and desiderata have arisen in the last 10 years, only some of which have been implemented. In this way, I evaluate the development of the CLARIN infrastructure (mainly its Netherlands part) over the past 10 years, and sketch the requirements and desiderata for the CLARIN infrastructure to address this research question for the next 10 years.

Keywords: text corpus search, treebank search, language acquisition, lexicon search, research infrastructure, CLARIN, CLARIAH

1 Introduction

Odijk (2011) sketched a research question on the acquisition of lexical properties of words, and illustrated it with some concrete examples, in particular with respect to the lexical properties of the Dutch synonyms *heel*, *erg*, and *zeer* (all meaning 'very'). This work also indicated what the CLARIN infrastructure should offer to make it possible to address this research question. Some of this research was actually carried out and reported on at various occasions (inter alia, Odijk

Acknowledgements: I would like to thank colleagues who commented on parts of earlier versions of this chapter, in particular Katrien Depuydt, Jesse de Does, Jan Niestadt, and Vincent Vandeghinste (all from the Institute for the Dutch Language) as well as anonymous reviewers of an earlier version of this chapter.

Jan Odijk, UiL-OTS, Utrecht University, Utrecht, the Netherlands, e-mail: j.odijk@uu.nl

https://doi.org/10.1515/9783110767377-028

2015, 2016, 2020a). When carrying out this research, new requirements and desirable features emerged, some of which were actually implemented.

Though the research question addressed was quite specific, the requirements to address this research question were formulated broadly, so that meeting these requirements enables many other linguistic research questions. Furthermore, the study of the acquisition of a linguistic property by children requires that one knows what the relevant facts of the adult language are, and it requires that one has a theory (model, grammar) of the adult I-language. So this research question also requires facilities to investigate the language of adults. For all of these reasons, it is interesting to investigate to what extent these requirements have actually been met.

In this contribution I sketch to what extent the CLARIN infrastructure has achieved these requirements and desiderata. The resulting picture is mixed: (1) some have been implemented; (2) some have not been implemented and are still highly desirable; (3) some have not been implemented but turned out to be not so urgent; (4) new requirements and desiderata have arisen in the last 10 years, only some of which have been implemented. In this way, I evaluate the development of the CLARIN infrastructure (mainly its Netherlands part) over the past 10 years, and sketch the requirements and desiderata for the CLARIN infrastructure to address this research question for the next 10 years.

I briefly sketch the original research problem in Section 2, introduce the requirements and desiderata derived from this research question in Section 3, and I evaluate their realization in the CLARIN infrastructure in three sections: Section 4 on searching in metadata, Section 5 on searching in lexicons, and Section 6 on searching in annotated corpora. I list new requirements that arose in the past 10 years in Section 7, and conclude this work in Section 8.

2 The research problem

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

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

(1) a. Hij is daar heel / erg / zeer blij over
he is there very / very / very glad about'
'He is very happy about that'
b. Hij is daar *heel / erg / zeer in zijn sas mee
he is there very / very / very in his lock with
'He is very happy about that'
c. Dat verbaast mij *heel / erg / zeer
That surprises me very / very / very
'That surprises me very much'

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

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

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

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

The central problem with regard to these data is now: how do children acquire these properties, in particular that *heel* does **not** take verbs and adpositions as modifiers while *erg* and *zeer* do.

2 Or maybe the whole VP *verbaast mij*.

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

3 Requirements

In order to address the research question formulated in Section 2, Odijk (2011) formulated a whole range of requirements that the CLARIN research infrastructure should meet. These requirements concern software and data. We list the requirements for software in Appendix A and the requirements for data in Appendix B.

The software requirements mostly concern options for searching, in metadata and in data. The data requirements list a number of corpora and lexicons that should be accessible and easily searchable.

In this chapter we assess to what extent CLARIN meets these requirements in 2021. We do so in three sections: one on metadata search (Section 4), one on lexicon search (Section 5), and one on corpus search (Section 6).

4 Metadata search

We first consider requirements that relate to search in metadata, as a first step towards identifying relevant data and selecting the ones needed for the research.

4.1 Realized

The requirement "Give me a list of all LRs for the Dutch language" is largely met by CLARIN. A simple query[4] in CLARIN's Virtual Language Observatory[5] yields many results (108,874 on 12 May 2021). This large number of resources is of course too large to inspect fully manually, and doing so would also not be very useful, because over 90,000 of the entries are titles of individual songs from the Dutch song database, as can be seen through this query.[6] The metadata are not at the right level of granularity for our purposes, so we carry out some further filtering. If we in addition select *resource type=corpus* we get a list of 134 corpora, still a long list but one that can be handled by a human. I filter further by selecting all options for modality except *modality=speech* using this query,[7] which leaves

4 https://vlo.clarin.eu/?fqType=languageCode:or&fq=languageCode:code:nld

5 https://vlo.clarin.eu

6 https://vlo.clarin.eu/search?q=liederenbank&fqType=languageCode:or&fq=languageCode:code:nld

7 https://vlo.clarin.eu/search?fqType=languageCode:or&fq=languageCode:code:nld&fqType=resourceClass:or&fq=resourceClass:corpus&fqType=modality:or&fq=modality:written&fq=modality:writtenlanguage&fq=modality:spoken

50 corpora. Not all these corpora are relevant for my research, so I would like to select the ones that are and store their description. This is in principle possible by making a virtual collection of the search result, and then removing the corpora that are not relevant from this virtual collection, and I succeeded in saving the query results as a virtual collection.[8]

I had to remove 3 corpora that did not validate. The remaining 47[9] indeed contain corpora that I was looking for, such as the Spoken Dutch Corpus, and the SoNaR corpus, and many others that are potentially relevant (e.g., the Dutch Parallel Corpus, Europarl data), some that are very relevant (e.g., the Basiscript and Basilex Corpora), but also some that are obviously not relevant (e.g., corpora for Middle Dutch). The highly relevant Dutch CHILDES corpora, however, are unfortunately not contained in the search results, because they are not characterized as resourceClass=corpus.

4.2 Not realized

Requirement 2 "What is the size of all Dutch text corpora (in #tokens") has not been realized. This requirement may appear a very simple requirement and easy to realize. It is not completely trivial, because different measures are relevant for different resources and different research purposes, so each researcher who provides data may provide his own metric. Examples of such different metrics are token count, the number of documents, the number of turns taken (in a dialogue), and so on. In addition, many resources have overlap with other resources, or are derivatives of other resources (e.g., the original text of a different resource enriched with linguistic annotations), which complicates the problem considerably. But the main reason why this has not been realized is because there has not been any central coordination for this aspect of the metadata. CLARIN promotes CMDI as the framework for creating metadata (Broeder et al. 2010; Windhouwer and Goosen 2022). CMDI allows researchers to define their own metadata schemata so that there is a lot of flexibility, which was needed in the early years of

8 However, the system works with a web interface, and it shows many of the bad features that are unfortunately common for most web interfaces (for an overview, see Odijk (2018)). For example, one cannot save before all entries are validated (there should be a distinction between saving (possibly with errors) and submitting (with validation). The *Save* button is not in a fixed menu as in a decent interface, but at the bottom of the list of 50 resource descriptions (which keeps one scrolling all the time). And there are many other missing or less fortunate options, which I reported to the developers.

9 Unfortunately, publishing the virtual collection failed, so it is a private collection.

CLARIN because no one had a good overview of what metadata were needed for the available language resources. But there were hardly any minimum requirements on which metadata information must be specified and how it should be specified. As a consequence, when all these metadata were brought together and made accessible via the VLO, the result turned out to be quite messy. This was observed by many, and Odijk (2014) carried out a detailed analysis of the problems and made several suggestions for improvements. The situation has significantly improved since then by the CLARIN CMDI Taskforce,[10] by the CLARIN Curation Task force (Ostojic, Sugimoto, and Ďurčo 2017), by the initiative on the CLARIN resource families[11] (Fišer, Lenardič, and Erjavec 2018; Lenardič and Fišer 2022), and by others, but is still not optimal.

A more complex query such as "Give me a list of all Dutch data that contain children between two and seven years old as speaker" is also not possible at this moment.

A query such as "Give me a list of all Dutch data containing any of the words *heel*, *zeer*, *erg*" is feasible via CLARIN's Federated Content Search (FCS),[12] but too few Dutch corpora currently have endpoints for FCS to make this useful.

5 Lexicon search

The requirement to find words that are closely related to *heel*, *erg* and *zeer*, for example adverbs that function as an intensifier ("booster") and that are synonymous or co-hyponyms of these words can be done via Cornetto (Vossen et al. 2013), for which a completely new search application was developed in CLARIN. For example, this query[13] searches for synonyms and co-hyponyms of the word *heel* as an adverb.

Cornetto includes the RBN dictionary (van der Vliet 2007), so search in RBN is also possible. Search in other dictionaries containing synonym or synonym-like information was therefore not needed (puzzle dictionaries were suggested in (Odijk 2011) as a backup alternative).

10 https://www.clarin.eu/sites/default/files/clarin2019_bazaar_nolda.pdf
11 https://www.clarin.eu/resource-families
12 https://www.clarin.eu/content/federated-content-search-clarin-fcs
13 http://cornetto.clarin.inl.nl/simple_search.xql?type=LE&purpose=S&id=d_r-106880

6 Search in annotated corpora

Many requirements involve search in annotated corpora. Many corpora have been annotated mostly at the token level, that is, linguistic properties are assigned to tokens. In some corpora, utterances have been enriched with syntactic structures. Such annotated corpora are called treebanks.

Many words in natural language are ambiguous, and this is also true of *heel*, *erg*, and *zeer*. In fact, each one is multiply ambiguous. We should be able to search for these words under the intended interpretation. The ambiguity is eliminated or significantly reduced by knowing the syntactic context of these words. Treebanks can be used to achieve this to a high degree, so we should be able to search in treebanks. I started my research using a corpus of CHILDES data in a search application that was created for a completely different research question (COAVA, (Cornips et al. 2016)). This corpus did not contain syntactic structures (it was not a treebank), and if I had based my research solely on this corpus I would have reached wrong conclusions. For details see Odijk (2016: 53). A treebank is required for this research.

A user-friendly treebank search application was developed outside the context of but clearly inspired by CLARIN: LASSY Word Relations Search (Tjong Kim Sang, Bouma, and van Noord 2010). After running for a few years it was not really maintained systematically, was regularly down and there was a real danger that it would disappear. In the context of CLARIN an update of this application was made, resulting in PaQu (Odijk et al. 2017). PaQu has been used extensively for addressing the research question, and it was especially suited for this because it has special provisions for searching for word relations, a crucial property for investigating the modification potential of words and its acquisition.

In the context of the cooperation between the Netherlands and Flanders on CLARIN, a new treebank search application was developed with query-by-example as its main distinguishing feature: GrETEL (Augustinus, Vandeghinste, and Eynde 2012; Augustinus et al. 2017). This application has also been used a lot for this and other research, and several improved versions of the application have been created (e.g., Odijk, van der Klis, and Spoel 2018).

These applications offer a number of treebanks for search, but they also allow a user to upload the user's own corpus, which is then parsed resulting in a treebank, which is then available for search. This feature has turned out to be very useful, and it made it possible to turn data for which no treebank existed into a treebank. It thus also enabled searching in treebanks derived from CHILDES corpora (which was one of the requirements), and a treebank for the Dutch CHILDES corpora was made generally available in PaQu.

Queries such as

1. find me sentences containing occurrences of the lemma *erg* of any part of speech (POS) which acts as a modifier to another word of any POS;
2. for each child, give a list of pairs (session, age) of the child;
3. for each child, give me #sessions by period, where period is e.g., every month, week, half year, year;
4. for child and each session, give #occurrences of *zeer*, *heel*, *erg*;

can be carried out. Others, which require more advanced aggregation of data currently cannot be carried out when using the applications mentioned:

1. for each child give me the list of new words uttered by period;
2. for child and each session, give #occurrences of *zeer*, *heel*, *erg*, by period;
3. give me utterances containing occurrences of *zeer*, *erg*, *heel* uttered by the child before any adult used any of these words;
4. give me #occurrences of *heel* uttered by the parent before the child utters it (idem for *zeer*, *erg*, etc.);

These have to be carried out by exporting the search results and do the analysis with different software. Exporting search results is possible, though there are severe limitations due to IPR. Therefore it is necessary to be able to carry out such queries and analyses inside the application.

For token-annotated corpora several search applications have been created, in particular the OpenSoNaR application (van de Camp, Reynaert, and Oostdijk 2017; de Does, Niestadt, and Depuydt 2017), which not only gives access to the 550 million token SoNaR corpus (Oostdijk et al. 2013) but also to the Spoken Dutch Corpus (CGN, (Oostdijk et al. 2002)), including its audio. And several search applications have been made available outside the context of but in close collaboration with CLARIN. These include search applications for modern Dutch (e.g., CHN (Contemporary Dutch Corpus)), but also for historical varieties of Dutch (e.g., Corpus Gysseling, Nederlab (Brouwer, Brugman, and Kemps-Snijders 2016; Brugman et al. 2016))

We discuss the current status of some other requirements:

– All annotated corpora contain errors. This is true not only for automatically annotated corpora but also for manually annotated corpora. None of the search applications have systematic provisions for reporting such errors. Reporting such errors so far goes via e-mail, which is not an ideal situation.
– Support for batch processing of queries is explicitly supported by OpenSoNaR. In PaQu and GrETEL one can achieve similar results by a combination of alter-

natives in a single query, made easier by using macros, in combination with the options for analysing the search results.
- All search applications can combine metadata and content search, but each does it in a different way, and all have limitations.
- In OpenSoNaR and the treebank applications one can formulate queries such as:
  1. give absolute and relative frequencies of *heel*/*hele*/*erg*/*erge*/*zeer* as adj by text genre, and speaker/participants education level, and by corpus;
  2. idem but for the word + the following POS-tag;
  3. idem but in the fully parsed part of CGN and in LASSY + the POS-tag of the modifiee head;
- To my knowledge, the requirements in (9) of Appendix A, i.e that new data created by enriching existing data is dealt with fully automatically in a fully CLARIN-compatible way has not been realized anywhere within the Netherlands and perhaps not even in Europe.
- Concerning the requirement (10) of Appendix A, i.e. maximizing the use of restricted vocabularies with well-defined semantics, a lot of work has been done on it, but in my view it is still insufficient to ensure true interoperability. The systems to store the vocabularies and their semantics changed over time (initially ISOCAT, since 2015 the CLARIN Concept Registry, and a new change is immanent). They usually had other uses by other communities as well, which often complicated things, and none of these systems had their concepts organized in such a way that it was easier to reuse existing ones than creating new ones. This topic is too broad to deal with properly here, so I will leave it at these general remarks.

7 New requirements

During our research, we found that we need many new features of the treebank query applications. Many of these were described in Odijk (2020b).

All annotated corpora contain errors. If one wants to draw reliable conclusions on the basis of corpus data, one has to assess the quality of the annotations in the corpus. In most cases a full manual evaluation is not feasible since the amount of data is too large. In those cases one can evaluate a representative sample of the data. But the treebank search applications should support selecting such representative samples. Currently PaQu offers some support for this (only via the word relations interface), but it is lacking in GrETEL and OpenSoNaR.

One technique that is especially effective for investigating recall of a search query is to formulate a query that searches for (as small as possible) a superset of the query results. For example, a treebank search for two verbs in a particular syntactic configuration can be generalized to a search for two verbs in any syntactic configuration (Bloem 2016). Formulating such a query can be quite difficult (see Odijk 2020b: 32–33). It would be good if the search applications would provide support for this, for example, by automatically suggesting the relevant queries on the basis of the original query.

One should also have the opportunity to annotate utterances in the search results, or specific words or phrases in search results to mark errors in the annotation or add information that is not present in the corpus (e.g., semantic information in a treebank). Ideally one would be supported in this by lookup in or even bootstrapping from external lexical resources (e.g. the CELEX lexicon (Baayen, Piepenbrock, and Gulikers 1996), Cornetto, or the Open Dutch Wordnet (Postma et al. 2016)). And it should of course be possible to use such annotations in the analysis component of the search application. Experiments with combining corpus search with search in external lexical resources have been done under the name "Chaining Search" (Dekker, Fanee, and de Does 2019), but the results of these experiments have not been integrated in any of the search applications.

Extensions of the analysis components (even the most advanced one, that found in GrETEL) are also desirable. The analysis component of GrETEL enables a user to combine arbitrary attributes of nodes that match with node descriptions in the query and metadata in a pivot table. But one should also be able to include computed relations between nodes, such as "node1 precedes / follows/ contains / overlaps with node2", "node1 is adjacent to node2", or "node1 and node2 are in a projective / non-projective grammatical relation",[14] as well as user definable ranges of numerical and date values.[15] Ideally, for advanced users a full database query language with functionality comparable to that of SQL would be provided,[16] but currently that is certainly not the case.[17]

An important feature of an analysis component is that one can easily get from an aggregate (e.g., the frequency of the combination of a token property, a node property and/or a metadata property) to the actual examples on which this is based. This feature has been implemented very well and is efficient in PaQu and

14 That is, informally stated: the relation between two nodes is projective if there are no crossing branches in a phrase structure tree over the surface string.

15 A limited number of these is actually possible, but not in a very user-friendly way.

16 The XQuery language would be the natural candidate for PaQu and GrETEL.

17 Such functionality is offered by the Prague Mark-up Language Treebank Query (PMLTQ) system, https://lindat.mff.cuni.cz/services/pmltq/#!/home.

in OpenSoNaR, but it has more limitations and is very inefficient in GrETEL 4. Other search applications (e.g., Nederlab) have only very limited options here.

It is often necessary to execute one and the same query at multiple occasions or by different researchers. However, it is currently not possible to store queries in the application so that they can be reused, though this is clearly a desirable feature. Our experiences with facilities to store queries in other applications (e.g., in SHEBANQ[18]), taught us that it is also necessary to carefully organize the storage of queries in order to make them easily findable for reuse: a simple list of stored queries is not enough because this list tends to get quite large very soon.

We also found several times that we wanted to compare results of two queries. It is therefore desirable if results of queries can be stored and set-like operations (union, difference, intersection) can be applied to stored queries, as e.g. MIMORE offers (Barbiers et al. 2016).

Some problems are caused by the nature of the syntactic structures in the treebanks for Dutch (Odijk et al. 2017: Section 23.3). One problem with the *de facto* standard treebank format is that single words that form a phrase on their own are not dominated by a phrase node: so in *de man sliep* 'the man slept' there is a node labeled NP for the phrase *de man*, but in *Jan sliep* 'Jan slept' there is no node labeled NP for the (single-word) phrase *Jan*. This complicates almost all queries, as also observed by Van Eynde, Augustinus, and Vandeghinste (2016: 106–107). It is clearly desirable that for each treebank a version in which there are nodes for all single word phrases is made available. This is not difficult to achieve since the relevant information to construct these phrasal nodes is present in the treebanks.

A second problem concerns so-called index-nodes. If a word or phrase has multiple functions in an utterance, the syntactic structure for this utterance contains multiple nodes for it: apart from the node that one expects (which we will call the antecedent), one or more nodes may occur that contain only an index and a grammatical relation as properties and that are coindexed with the antecedent. Other properties of their antecedent are not present at this node. It is very difficult to define queries in Xpath to obtain all properties of the antecedent of an index node.[19] It is desirable to provide a version of the treebanks in which such index nodes are replaced by a copy of their antecedents. This feature actually has recently been implemented in PaQu,[20] but it is not available in GrETEL.

Finally, it should be possible for a user to share corpora uploaded by him/her with a group of selectable users. Currently, some applications either keep

18 See https://shebanq.ancient-data.org/hebrew/queries.

19 See the DACT Cookbook, Section Antecedents of co-indexed nodes for an implementation of inclusion of indexed nodes in Xpath.

20 https://paqu.let.rug.nl:8068/info.html#expanded

an uploaded corpus private to the user, or make it openly available to all users. This is a problem because a user does not want to bother everybody with his/her uploaded corpora (e.g., in an experimental phase), and because a user may want to share the data only with a small group of collaborators during the initial phase of a research project.

8 Conclusions

I sketched to what extent the CLARIN infrastructure has achieved requirements and desiderata put forward by Odijk (2011) on the basis of a research question. The resulting picture is mixed: (1) some have been implemented; (2) some have not been implemented and are still highly desirable; (3) some have not been implemented but turned out to be not so urgent; (4) new requirements and desiderata have arisen in the last 10 years, only some of which have been implemented. In this way, I evaluated the development of the CLARIN infrastructure (mainly its Netherlands part) over the past 10 years, and gave a sketch of the requirements and desiderata for the CLARIN infrastructure to address this research question (and many others) in the next 10 years. It is my hope that these new requirements and desiderata will be taken up in future projects both at the ERIC level (where appropriate) and at the national level.

Appendix A: Software requirements

1. Give me a list of all LRs for the Dutch language.
2. What is the size of all Dutch text corpora (in #tokens)?
3. Give me a list of all Dutch data that contain children between two and seven years old as speaker.
4. Give me a list of all Dutch data containing any of the words *heel*, *zeer*, *erg*.
5. Find words that are closely related to *heel*, *erg*, and *zeer*, e.g., adverbs that function as an intensifier ("booster") and that are synonymous or co-hyponyms. A recursive search for synonyms is therefore desirable, limited by a maximum depth (since otherwise there is no guarantee the process will finish), and for each found synonym the level of depth at which it was found. The search engine should be clever enough to determine that this kind of information can be found in (certain) dictionaries, but not, e.g., in text or speech corpora, preferably without having to search through all these data (e.g. based on metadata, or based on a classification of types of resources).

6. As with many words in natural language, each of the three words is multiply ambiguous, so we should be able to search for these words under the intended interpretation.
7. Treebanks can achieve this to a high degree, so we should be able to search in treebanks.
   (a) Queries such as:
      i. Find me sentences containing occurrences of the lemma *erg* of any POS which acts as a modifier to another word of any POS.
      ii. For each child, give list of pairs session + age of the child
      iii. For each child, give me #sessions by period, where period is e.g., every month, week, half year, year.
      iv. For each child give me the list of new words uttered by period.
      v. For child and each session, give #occurrences of *zeer*, *heel*, *erg*.
      vi. Idem, by period.
      vii. Give me utterances containing occurrences of *zeer*, *erg*, *heel* uttered by the child before any adult used any of these words.
      viii. Give me #occurrences of *heel* uttered by the parent before the child utters it (idem for *zeer*, *erg*, etc.).
   (b) Treebanks contain errors. I would like to report the errors I found in the treebank in a systematic manner (so provisions for that should be available).
   (c) Batch processing of queries should be supported, or there should be a simple way of issuing the same query for different lexical items without too much manual work. (e.g., a map function that applies a query to each item in a list of lexical items, and yields a list of query results per lexical item).
   (d) Some simple queries use a mix of metadata and content search, and the content search is on multiple tiers, so that should be possible in the search engine
   (e) In the CHILDES corpus, we again run into the problem of the ambiguity of the words. So perhaps I would like to parse these corpora (or at least the parts where adults speak),
8. POS-tagged corpora such as CGN and SoNaR can also be useful and are usually larger than treebanks. We would like to be able to formulate queries such as:
   (a) Give absolute and relative frequencies of *heel*/*hele*/*erg*/*erge*/*zeer* as adj by text genre, and speaker/participants education level, and by corpus.
   (b) Idem but for the word + the following POS-tag.
   (c) Idem but in the fully parsed part of CGN and in LASSY + the POS-tag of the modifiee head.

9. Of course, the found and newly created data
   - should be stored in a supported format;
   - with automatically generated metadata;
   - with automatically generated provenance data;
   - using data categories mapped to or from ISOCAT;
   - for which PIDs are provided;
   - stored on a server of a CLARIN-centre;
   - so that they can become proper resources on their own;
   - and are visible, accessible and interpretable as part of enriched publications
10. Even simple and well-definable data categories at the time allowed any string as value. These should be defined in a very strict manner, at least by specifying a regular expression for the values they can take. If any string can be filled in, no search engine can do anything with it that makes sense.

Appendix B: Data requirements

1. Dutch EuroWordnet (in 2011 it was only available as a download via ELRA M0016).
2. Or Cornetto (in 2011 available as a download via the Dutch HLT-Agency).
3. Ordinary dictionaries containing synonyms (e.g., Van Dale dictionaries, perhaps RBN).
4. Puzzle dictionaries with synonym information.
5. Relevant data can be found in the CHILDES system (part of TalkBank), with 7 corpora for Dutch, but of course with their own data formats (CHAT) and tools (CLAN).
6. Spoken Dutch Corpus.
7. SoNaR Corpus.

Bibliography

Augustinus, Liesbeth, Vincent Vandeghinste & 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 & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Augustinus, Liesbeth, Vincent Vandeghinste, Ineke Schuurman & Frank Van Eynde. 2017. Gretel: A tool for example-based treebank mining. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 269–280. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.22. License: CC-BY 4.0.

Baayen, R H., R Piepenbrock & L. Gulikers. 1996. *Celex2*. Philadelphia: Linguistic Data Consortium. LDC96L14: https://catalog.ldc.upenn.edu/LDC96L14.

Barbiers, Sjef, Marjo van Koppen, Hans Bennis & Norbert Corver. 2016. MIcrocomparative MOrphosyntactic REsearch (MIMORE): Mapping partial grammars of Flemish, Brabantish and Dutch. *Lingua* 178: 5 – 31. Linguistic Research in the CLARIN Infrastructure.

Bloem, Jelke. 2016. Evaluating automatically annotated treebanks for linguistic research. *Proceedings of the 4th workshop on challenges in the management of large corpora (CMLC-4)*, 8–14. Portorož, Slovenia: European Language Resources Association (ELRA).

Broeder, D., M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg & C. Zinn. 2010. A data category registry- and component-based metadata framework. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC 2010)*, 43–47. Valetta, Malta: European Language Resources Association (ELRA).

Brouwer, Matthijs, Hennie Brugman & Marc Kemps-Snijders. 2016. A Solr/Lucene based multi tier annotation search solution. *Selected papers from the CLARIN annual conference 2016, 26–28 October, Aix-en-Provence*, 29–37. Linköping, Sweden: Linköping University Electronic Press.

Brugman, Hennie, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang & Antal van den Bosch. 2016. Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Camp, Matje van de, Martin Reynaert & Nelleke Oostdijk. 2017. WhiteLab 2.0: A web interface for corpus exploitation. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 231–243. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.19. License: CC-BY 4.0.

Cornips, Leonie, Jos Swanenberg, Wilbert Heeringa & Folkert de Vriend. 2016. The relationship between first language acquisition and dialect variation: Linking resources from distinct disciplines in a CLARIN-NL project. *Lingua* 178: 32 – 45. Linguistic Research in the CLARIN Infrastructure.

Dekker, Peter, Mathieu Fanee & Jesse de Does. 2019. CLARIAH chaining search: A platform for combined exploitation of multiple linguistic resources. In K. Simov & M. Eskevich (eds.), *Proceedings of CLARIN annual conference 2019*, Theory and Applications of Natural Language Processing, 24–27. CLARIN.

Does, J. de, J. Niestadt & K. Depuydt. 2017. Creating research environments with BlackLab. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 245–257. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.20. License: CC-BY 4.0.

Fišer, Darja, Jakob Lenardič & Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the eleventh*

international conference on language resources and evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

Lenardič, Jakob & Darja Fišer. 2022. The CLARIN Resource and Tool Families. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Odijk, Jan. 2011. User scenario search. internal CLARIN-NL document, http://www.clarin.nl/sites/default/files/User %20scenario%20Serach%20110413.docx, last accessed 2022-03-25.

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

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

Odijk, Jan. 2016. A Use case for Linguistic Research on Dutch with CLARIN. In Koenraad De Smedt (ed.), *Selected papers from the CLARIN annual conference 2015, October 14–16, 2015, Wroclaw, Poland*, Linköping Electronic Conference Proceedings no. 123, 45–61. CLARIN, Linköping, Sweden: Linköping University Electronic Press. http://www.ep.liu.se/ecp/article.asp?issue=123&article=004, http://dspace.library.uu.nl/handle/1874/339492.

Odijk, Jan. 2018. Why I do not like web interfaces for data entry. Utrecht University, https://dspace.library.uu.nl/handle/1874/375225.

Odijk, Jan. 2020a. CLARIN-supported research on modification potential in Dutch first language acquisition. *Selected papers from the CLARIN annual conference 2019*, Volume 172 of *Linköping Electronic Conference Proceedings*, 94–107. Linköping, Sweden: Linköping University Press.

Odijk, Jan. 2020b. De verleidingen en gevaren van GrETEL. *Nederlandse Taalkunde* 25 (1): 7–38.

Odijk, Jan, Martijn van der Klis & Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. *Proceedings of the 16th international workshop on treebanks and linguistic theories (tlt16)*, 46–55. Prague, Czech Republic. http://aclweb.org/anthology/W/W17/W17-7608.pdf.

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

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman. 2013. The construction of a 500 million word reference corpus of contemporary written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for dutch: Results by the STEVIN-programme*, 219–247. Berlin: Springer. http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat & Harald Baayen. 2002. Experiences from the Spoken Dutch Corpus project. *Proceedings of the third international conference on language resources and evaluation (LREC-2002)*, 340–347. Las Palmas: ELRA.

Ostojic, Davor, Go Sugimoto & Matej Ďurčo. 2017. The Curation Module and Statistical Analysis on VLO Metadata Quality. In Lars Borin (ed.), *Selected papers from the CLARIN annual conference 2016, Aix-en-Provence, 26–28 October 2016*, Linköping Electronic Conference Proceedings no. 136, 90–101. CLARIN, Linköping, Sweden: Linköping University Electronic Press. https://ep.liu.se/ecp/article.asp?issue=136&article=007&volume=0#.

Postma, Marten, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen & Piek Vossen. 2016. Open Dutch WordNet. *Proceedings of the eighth global WordNet conference*. Bucharest, Romania.

Tjong Kim Sang, Erik, Gosse Bouma & Gertjan van Noord. 2010. LASSY for beginners. Presentation at CLIN 2010, Utrecht, http://ifarm.nl/erikt/talks/clin2010.pdf, last accessed 2022-03-25.

Van Eynde, Frank, Liesbeth Augustinus & Vincent Vandeghinste. 2016. Number agreement in copular constructions: A treebank-based investigation. *Lingua* 178: 104 – 126. Linguistic Research in the CLARIN Infrastructure.

Vliet, H.D. van der. 2007. The referentiebestand Nederlands as a multi-purpose lexical database. *International Journal of Lexicography* 20 (3): 239–257.

Vossen, Piek, Isa Maks, Roxanne Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang & Maarten de Rijke. 2013. Cornetto: a lexical semantic database for Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for dutch, results by the STEVIN-programme*, Theory and Applications of Natural Language Processing, 165–184. Berlin Heidelberg: Springer.

Windhouwer, Menzo & Twan Goosen. 2022. Component Metadata Infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.