**User Scenario Search in CLARIN**

***Jan Odijk 110413***

**There is also a presentation about more or less the same topic.**

I am interested in the problem of language acquisition, more specifically first language acquisition by young children learning their native language.

The basic research questions are questions such as:

* What is inborn in young children enabling them to acquire natural language so effortlessly
* Which aspects are learned on the basis of input, i.e. not inborn
* How do inborn aspects and learning interact?
* What is the acquisition speed
* Can any stages be distinguished in the acquisition process, e.g. are certain aspects of language acquired earlier than others; do innate capabilities develop in different stages, if any? Etc.

The concrete problem I would like to address to get some insight into such questions revolves around the acquisition of syntactic combinatorial potential of lexical items.

The first example concerns the (Dutch) words *zeer, heel*, and *erg*.

These words are synonyms in Dutch and they mean the same as English *very*. All these words can modify adjectives as an intensifier:

1. Hij is heel/erg/zeer ziek

Dutch also allows prepositional phrases (PP) as predicates. Examples are *in zijn nopjes met ‘happy with’, in zijn sas ‘happy’, in verwachting ‘pregnant’.* When such PPs have a gradable meaning, they can also be modified by *erg* and *zeer:*

1. Hij is zeer/erg in zijn nopjes met dat voorstel

However, they cannot be modified by *heel*:

1. \*Hij is heel in zijn nopjes met dat voorstel

The difference between *zeer, erg* on the one hand and *heel* on the other cannot be ascribed to a meaning difference between these words, since they are perfect synonyms. It can also not be ascribed to a meaning difference between adjectival predicates and PP predicates, since with predicates with identical or very similar meaning the same difference occurs:

1. Blij v. in zijn nopjes
	1. Hij is erg / zeer / \*heel in zijn nopjes met dat voorstel
	2. Hij is erg / zeer / heel blij met dat voorstel
2. Tevreden v. tegen
	1. Hij is erg / zeer / heel tevreden over dat voorstel
	2. We zijn erg / zeer / \*heel tegen (dat voorstel)
3. Inde war v. verward
	1. Hij is erg /zeer / \*heel in de war
	2. Hij is erg / zeer / heel verward
	3. Hij gedraagt zich erg / zeer / heel verward
4. Zwanger v. in verwachting
	1. Zij is erg/zeer/heel zwanger
	2. Zij is erg/?zeer/\*heel in verwachting.

We see a similar phenomenon in English *very*. It can modify adjectives, but not verbs or PPs:

1. English:
	1. He is very angry
	2. \*He loves her very
	3. \*he is very in love

The verb *love* is gradable, but if intensified it has to be modified by *very much* rather than by *very*.

I conclude that the difference between *zeer/erg* and *heel* is a purely syntactic difference: *zeer/erg* can modify gradable adjectival, verbal, and PP predicates, while *heel* can modify only gradable adjectival predicates. Similarly, for English, *very* can modify adjectives, but not verbs or PPs.

A second example involves the Dutch verbs *worden* and *raken* (in its copular sense), which are almost synonymous, corresponding approximately to English *become* and *get* (in its copular sense). As copular verbs, they can be combined with a subject and a predicate, e.g. an adjectival predicate:

1. Worden v. raken
	1. Zij werd zwanger ‘she became pregnant’
	2. Zij raakte zwanger ‘she got pregnant’

The difference in meaning between the two verbs is very small, basically *worden* is neutral wrt the accidental character of the subject-predicate relation, while *raken* implies the accidental character of the subject predicate relation. There are two differences between these verbs, related to the syntactic category of the predicate. First, *raken* allows PP predicates, while *worden* does not:

1. PPs
	1. Zij raakte in verwachting ‘she got pregnant’
	2. \*Zij werd in verwachting ‘she became pregnant’

Second, *worden* allows NP predicates, but *raken* does not:

1. NPs
	1. Zij werd moeder ‘she became a mother’
	2. #Zij raakte moeder ‘she got a mother’

Again, the difference cannot be ascribed to semantic properties: after all the PP expression ‘in verwachting is synonymous to the AP ‘zwanger’.

It is also clear that the properties of these words cannot be innate, or be derived from other facts, but must instead be learned as idiosyncratic properties of these words. One may assume that the properties of *heel* en *erg* can be learned from exposure to examples of these words modifying adjectival and PP predicates. Also, the fact that *heel* and *very* can modify adjectival predicates can be learned from exposure to examples of these words modifying adjectival predicates. But how can a native speaker of Dutch learn that *heel* cannot modify PP predicates, and how can a native speaker of English learn that *very* cannot modify verbs?

Similarly, the fact that *worden* can take an NP or AP predicate, and that *raken* can take a PP or AP predicate can be learned from exposure to examples of these words taking such predicates. But how can a native speaker of Dutch learn that *worden* cannot combine with PP predicates, and that *raken* cannot combine with NP predicates?

It is generally assumed that a native speaker is not presented with negative evidence, i.e. examples of strings of which it is made explicit that they are not well-formed, and even if they are, such statements usually do not have much influence, as can be witnessed from the fact that phenomena that are ill-formed according to the official norms of a language and against which people are warned, are nevertheless used.

Many hypotheses on how to explain these phenomena have been proposed and have partially been investigated. I just mention the names of some of these hypotheses:

1. Possible relevant factors towards an explanation
	1. Indirect Negative Evidence Hypothesis
	2. Semantic Bootstrapping Hypothesis
	3. Stage of Acquisition hypothesis
	4. Sensitive acquisition period for each lexical item (or classes of lexical items) to acquire its syntactic properties

Obviously, we will also want to investigate whether we can correlate the facts mentioned to independent properties of the relevant lexical items, such as

* Is there a relation with the fact that *zeer* appears to be rather formal, while *heel* and *erg* are not?
* Is there a relation with adverb-adjective agreement (substandard):
	+ heel/*hele dikke boeken ‘very thick books’*
	+ *erg/erge dikke boeken*
	+ *Zeer/\*zere dikke boeken*

And we do not want to restrict ourselves to the specific lexical items mentioned, but also want to investigate how other, closely related words behave in these respects.

In order to test such hypotheses (or combinations of them), we need empirical data, and here easy search in complex existing data comes into play

This leads to the following queries that (hopefully) will provide us with the data that we need to answer the questions posed, and we want to be able, in the CLARIN context, to find all relevant data from one interface, and without even knowing what data there are and where they are.

So obvious queries would be

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 2-7 years old as speaker
4. Give me a list of all Dutch data containing any of the words *heel*, *zeer, erg*

Most of these (pretty simple) queries can currently not be posed at all from a single place, even if we restrict attention to the data hosted at one centre (e.g. INL, or MI).

CLARIN should change this situation!

And of course, **I should be able to store this result somehow as a virtual collection**.

Next, I would like to consult existing grammatical descriptions for such phenomena, e.g. by searching (in a clever way) in the Taalportaal. However, since the Taalportaal does not exist yet, this cannot be done yet and I could not experiment yet; But it will impose restrictions on how the Taalportaal should be structured and how it can be searched.

Another query that would be interesting is:

1. Find words that are closely related, e.g. adverbs that function as an intensifier (‘booster’) and that are synonymous or co-hyponyms

Using e.g.

* Dutch [EuroWordnet](http://www.illc.uva.nl/EuroWordNet/) (currently via [ELRA](http://catalog.elra.info/index.php) [M0016](http://catalog.elra.info/product_info.php?products_id=545))
* Or [Cornetto](http://www.inl.nl/nl/lexica/cornetto) (via the Dutch [HLT-Agency](http://www.inl.nl/nl/tst-centrale))
* Ordinary dictionaries containing synonyms (e.g. Van Dale dictionaries, perhaps RBN
* Puzzle dictionaries with synonym information

In ordinary electronic dictionaries and puzzle dictionaries, we may want to find all synonyms, but typically **not all** synonyms are provided for a given word. If one looks for the synonyms of a word A, one will e.g. find {B,C}, but if one looks for synonyms for B, one might find e.g. {C,D} (i.e. without A, but with additional word D which was not mentioned as a synonym of A. **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**. Furthermore, manual intervention after each step should be possible, since many synonyms of a word A are only a synonym for word A in a sense that the researcher is not interested in.

I did such an experiment manually, for the words *zeer, heel* and *erg*, using Van Dale and the puzzle dictionary <http://puzzel.woordenboek.nu/woordenboek/synoniemen/> and came up with a list such as the following:

1. *abnormaal afschuwelijk akelig bijster bijzonder bovenmatig buitengemeen buitensporig danig donders eminent enorm exceptioneel extra extraordinair extreem fabelachtig fenomenaal geweldig gigantisch intens kolossaal merkwaardig mirakels onbeschrijfelijk ongelofelijk ongehoord ongekend ongemeen onmenselijk onmetelijk ontzettend onwijs speciaal uitermate uiterst uitzonderlijk verdraaid verduiveld verrekte verschrikkelijk vet zeldzaam …..*

(but I did not keep track of the recursion depth).

Carrying out such a search in Cornetto or EuroWordnet should give such results in an easier manner, and though I have not yet been able to experiment with that, I fear that the results will be disappointing unless a recursive search is allowed. See the Cornetto interface at <http://deb.fi.muni.cz/install.php/> (Ik krijg die niet geïnstalleerd (“Firefox kon het bestand op http://deb.fi.muni.cz/install.php/clients/debvisdic\_cornetto.xpi

niet installeren omdat: Geen geldig installatiepakket”, en dat is al drie jaar zo)

An interesting problem for such searches in the CLARIN context is that the user should not have to know which resources exist that might provide such information. **Thus, 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).**

Next, I would like to first check the basic factual assumptions I have made above about the syntactic properties of words such as *zeer, erg, heel*, and all their synonyms, and for *worden* and *raken*, and all their synonyms. **This means that the output of the synonym search should be easily used in a next query.**

If we do this for the words *zeer, erg* and *heel*, we encounter a problem, and this problem is not specific to these words but pervasive throughout natural language: each of the 3 words is multiply ambiguous!

* Erg (4x)= noun(de) ‘erg’; noun(het)’evil’, adj+adv ‘unpleasant’, adv ’very’
* Zeer (3x)= noun ‘pain’; adj ‘painful’; adv ‘very’
* Heel (3x) = adj ‘whole’; verbform ‘heal’; adv ‘very’

Fortunately, we have Pos-tagged corpora, so we might search for specific words provided they are tagged with a specific Pos-tag. This will reduce the problem somewhat, but for Dutch most corpora do not distinguish adj from adv by category!

Thus a search using Pos-tags will ensure that we will not get results with *zeer* and *erg* as a noun, but the other ambiguities cannot be avoided in this manner.

For this particular problem, one might achieve a reasonable approximations by looking for Pos-bigrams, e.g. the word *zeer* in the intended meaning will usually be immediately followed by and adj (if it modifies an adj), and by a VZ (if it modifies a PP). However, this does not solve the whole problem (VZ can follow the adj *zeer* even if the adj does not modify a PP), and for modification of verbs such a strict ordering does not even hold. For verbs, one might try to select a reasonable sample for *zeer* by requiring that a sentence contains *zeer* as an adj, that does not immediately precede an adj, and in which only one verb occurs. So the query is getting pretty complex (and still does not give exactly what the researcher needs), but it should be possible to formulate such a query.

However, there are more data, in particular there are text corpora in which each sentence is assigned a full parse tree (Treebanks), some manually verified (e.g. in a subpart of CGN, in LASSY-Small), some generated in a fully automatic manner (e.g. LASSY-Large).

A researcher can use such resources only in a sensible way if he/she knows what kind of information such resources contain, how they have been created, what the underlying grammatical theory is, how it I has been implemented (at a course level), so that he/she knows exactly what to ask for and what he/she can and cannot expect. That should be done in courses that explain such aspects of these corpora.

I did some experiments with queries involving the words *zeer, erg* and *heel*, using the (extremely simple) LASSY interface on <http://www.let.rug.nl/erikt/lassy/>. The query I wanted to use was not usable directly, because of peculiarities of the LASSY treebank, so I started with the queries:

1. *erg::mod:*
2. *zeer::mod:*
3. *heel::mod:*

The first one, e.g., means: **find me sentences containing occurrences of the lemma *erg* of any PoS which acts as a *modifier* to another word of any PoS.** The result is a list of sentences that satisfy these criteria, but what is more, it provides me with statistics on the PoS tags of *erg* and the PoS-tags of the modifies word (and many more statistics). From these statistics (which I can get only on the web page, using the simple interface mentioned), I constructed the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Modifiee**  | **erg**  | **zeer**  | **heel**  |
| ADJ  | 143  | 268  | 263  |
| WW  | 35  | 49  | **9**  |
| BW  | 1  | 1  | 7  |

i.e. *erg* modifies and ADJ 143 times, a WW (verb) 35 times, and a BW (adverb) 1 time, etc. I also found that with this query I only retrieve these words as ADJ (not as noun), and only once as a BW (which is an error in the LASSY Treebank)

And we notice that *heel* in the LASSY-SMALL corpus **does** modify verbs!

This requires a further investigation, and I issued the query

1. heel::mod:WW

 **i.e., find me all sentences in which *heel* occurs as any PoS modifying a WW (verb).** I analyzed the resulting set, and discovered that all examples (but one) involve modification of participles, some of which were real participles, and some of which should actually be analyzed as an adjective. Alpino actually makes this distinction internally, but the CGN-D-COI tag set apparently not (so each participle is analyzed as a verb, and that is all we get). In the remaining example, the LASSY parse tree contains an error.

In summary, I am confident that my basic facts are indeed correct, even after a check in the LASSY-Small Treebank, but **I would like to report the errors I found in the LASSY treebank in a systematic manner (so provisions for that should be available)**. The errors involve not only the one wrong parse I found, but also the fact that *heel* was assigned the PoS-tag BW once (it should be ADJ).

Of course, I would like to carry out such queries not only for these words, but also for all the synonyms and closely related words I found earlier, and I do not want to do that manually one by one. **So 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*).***

Above I searched for the *lemma heel, (erg, zeer, etc)*, but of course I am especially interested in the word forms *hele* and *erge* . The query

hele::mod:adj

yields 5 results.

For LASSY there is also a more sophisticated search interface called DACT, <http://github.com/danieldk/dact/downloads> , which has to be downloaded and installed locally. The data must be obtained from INL (free user license, but some handling costs), and also stored locally (I believe).

Next, I would like to investigate what examples of these lexical items occur in children’s speech, and in speech by adults addressed to children.

Relevant data can be found in the [CHILDES](http://childes.psy.cmu.edu/) system (part of [TalkBank](http://talkbank.org/)), with [7 corpora for Dutch](http://childes.psy.cmu.edu/data/Germanic/Dutch/), but of course with their own data formats ([CHAT](http://childes.psy.cmu.edu/manuals/chat.pdf)) and tools ([CLAN](http://childes.psy.cmu.edu/clan/))

Fortunately, these data are also mirrored at MPI and accessible via [(ANNEX/)TROVA](http://www.lat-mpi.eu/tools/annex). Daan Broeder did some experiments for me using TROVA. He actually searched not only in the CHILDES corpora but also in the CGN (with one query, if I understood well, which is good!). For children’s speech I am interested only in children between 2 and 7 years old. So Daan made a query with the following conditions

* Search in CGN and CHILDES
* Where the spoken language is Dutch
* And 2<age<7

(the exact details are in the slides (#40) but as a print out of the graphical user interface so difficult to see)

And a regular expression search on the result of this query (slide #41)

* heel|zeer|erg|erge|hele

which yielded too much, so he corrected it to

* regexp ^heel$|^erg$|…

(i.e. we are only interested in whole words *heel, erg*, etc)

And then he made an attempt to take PoS-tags also into account (for CGN only)

* regexp op WORDS tier + POS

This went reasonably, but the last query on CGN did not yield any result, which I could have told him in advance, since I know that CGN does not contain children younger than 15. The query system should, where possible, make use of such knowledge (which should be stored somehow in the metadata of CGN) to avoid unnecessary searching.

**Note that these 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**

I wanted the output to have the form of a CVS file with columns:

* + Corpus (e.g. Van Kampen Corpus)
	+ File: (e.g. laura74.cha)
	+ Line: (e.g. 139)
	+ Part Role: (e.g. Child)
	+ Child Gdr: e.g. female)
	+ Age: (e.g. 5;6.12)
	+ UTT (e.g. “ja , die s **erg** moeilijk .”)

And I am not sure that we can get this out in this format (the issue is not CVS, it may be XML, but the issue is about the headers). I started looking up the meaning (intended by me) for these labels in ISOCAT, but stopped because they were not easy to find.

One that I did find worried me very much. The DC **ParticipantAge** exists in ISOCAT (<http://www.isocat.org/datcat/DC-2550>), but its allowed values are any String!

There is also the DC ParticipantAgeRange <http://www.isocat.org/datcat/DC-3653>, buts possible values are also just String!

That is useless!!! Such simple and well-definable DCs 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.

(I reported this separately to Folkert and Menzo).

Anyway, results are e.g.

* *Heel* is found 153 times in Van Kampen corpus
* Erg is found 77 times in Van Kampen corpus
	+ But many are an irrelevant use of erg

Of course, 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), and TTNWW should offer such functionality, but whether that will help very much is as yet unclear to me.

Some other queries that I would like to be able to execute are:

* 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, in CGN and SONAR:

* Give absolute and relative frequencies of *heel*/hele/*erg/erge/*zeer as adj by text genre, and speaker/participants education level
	+ In CGN (spoken corpus)
	+ In SONAR (written corpus)
* Idem but for the word + the following Pos-tag
* Idem but in the fully parsed part of CGN and in LASSY + the PoS tag of the modifiee head

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
	+ Are visible, accessible and interpretable as part of enriched publications