Past and future of linguistics in the Netherlands

A very personal view

Jan Odijk
Utrecht University

This contribution sketches a very personal and subjective view on the past 50 years of linguistics in the Netherlands as well as on the next 50 years in linguistics in the Netherlands. I will be more specific on the past 10 years and on the next 10 years.

The past

Let me start by wholeheartedly congratulating the AVT (General Association for Linguistics) with the 50th anniversary of the Linguistics in the Netherlands conference (Taalkunde in Nederland, TIN).

The past 50 years

The past 50 years formed a very exciting period for linguists, in particular because during this period the most important and most influential linguist of the twentieth century and arguably of all times was active. Noam Chomsky had an enormous impact on linguistics in general (and far beyond), on linguistics in the Netherlands, and also on me personally. I probably would have become a linguist anyway, since the stories of my Greek teacher at high school about Greek nominal and verbal inflectional paradigms made me search for literature on historical linguistics and sound laws, but when (in 1972) I encountered a little booklet called 'Syntactic Structures' at the Rotterdam city library I immediately knew what I wanted to study.

I started my study in general linguistics in 1978, became a member of AVT in 1983, and had my first publication in the *Linguistics in the Netherlands* book in 1984. The 'TIN-dag' was a very important event for me as a young PhD-student.

https://doi.org/10.1075/avt.00020.odi
Linguistics in the Netherlands 36:1 (2019), pp. 43–48. issn 0929-7332 | e-issn 1569-9919

It offered (and still offers) an opportunity to present one's work in a friendly, informal and constructive environment, and contributes in this way to preparing PhD students for presenting their work to a large audience at peer-reviewed international conferences and workshops. The TIN-model is inclusive in nature and therefore also an excellent opportunity for networking. The model formed the inspiration for a similar conference in the area of *computational* linguistics, which was started in 1990 in Utrecht by Gertjan van Noord, Lisette Appelo and some other colleagues. This annual conference, called Computational Linguistics in the Netherlands (CLIN), has been a big success from the very start and is growing in size almost every year. In contrast to TIN, which is always held in Utrecht, CLIN is held in a different city each year, not only in the Netherlands but also (despite its name) in Flanders. CLIN celebrates its 30th birthday in 2020 in Utrecht, from where it started.

The past 10 years

My own interests have been in the areas of theoretical linguistics, especially syntax, and computational linguistics. The last 10 years I have been able to combine these in projects for large scale research infrastructures targeted at humanities researchers. The initial project (CLARIN-NL, which started in 2009) focused on all humanities researchers who work with natural language. It contributed to the international research infrastructure CLARIN. Linguistics has been a dominating discipline in the CLARIN-NL project. Later projects (CLARIAH-SEED and CLARIAH-CORE) did not have their focus exclusively on language and contributed not only to CLARIN but also to the DARIAH research infrastructure, but linguistics continued to be a major discipline for which infrastructural facilities were being created.

It has been our aim with this research infrastructure to enable linguists to base their research on large amounts of data, orders of magnitude larger than ever before (creating a larger empirical basis for theories and hypotheses), to use computational techniques to do this fast and efficiently, thus trying to accelerate the pace of linguistic research, and at the same time to do this in such a way that the learning curve is not too steep for linguistic researchers.

I want to highlight a small number of these infrastructural facilities that are particularly relevant to linguistics. I especially highlight facilities for the Dutch language, which has been the focus of the projects in the Netherlands, even though CLARIAH and CLARIN are of course not restricted to this language. They represent just a tiny fraction of the applications, services and data made available in the CLARIN infrastructure. For more extensive overviews I refer to

CLAPOP,[1] the CLARIN virtual language Observatory (VLO),[2] and to Odijk & Van Hessen (2017).

The first group of applications that I would like to highlight are applications for linguistically enriching text corpora. *Frog* enriches a Dutch text with tokenisation information, part of speech tags, lemmas, syntactic chunks, named entity types, and even some dependency parsing. The *Alpino* parser enriches each sentence in a Dutch text corpus with a syntactic structure. These applications have been incorporated into the CLARIAH infrastructure in such a way that, I claim, any linguist can apply them to his/her own text corpus.

The second group of applications that I would like to highlight concern applications for searching and analysing linguistically enriched corpora. *MIMORE* enables combined search and analysis in three different databases with dialectal data of Dutch. *OpenSoNaR(+)* enables search in corpora that are linguistically enriched at the token level, currently in (the 550 million token) *SoNaR* written corpus and the (10 million token) *Spoken Dutch Corpus (CGN)*. Indexing techniques make it possible to search through such large amounts of data very efficiently.

AutoSearch allows you to upload your own linguistically enriched text corpus. The treebank applications *PaQu* and *GrETEL 4* enable you to upload a text corpus in various formats (plain text, CHAT, TEI, FoLiA), and have it automatically linguistically enriched by Alpino, after which the corpus is available for search and analysis.

All these search and analysis applications have multiple interfaces for formulating queries, ranging from beginner level to expert level, many with levels in between. This makes it very easy for a linguist to start using these applications and gradually build up the expertise for full linguistic analyses using these facilities.

The future

It is sometimes said that it is difficult to make correct predictions, especially if they concern the future.[3] But this is false. Predictions are made about the unknown, and these can be in the future or the past. We can influence, even shape the future, at least to some degree (but not the past). So, though it remains true that it is difficult to make correct predictions, it is actually a little bit easier when they concern

1. http://portal.clarin.nl/clariah-tools-fs

2. https://vlo.clarin.eu/

3. See https://quoteinvestigator.com/2013/10/20/no-predict/ for an investigation into who might be the source of this statement (or equivalent variants).

the future. I will sketch here how we try to shape the future in the next ten years by briefly outlining some of the plans that will be carried out in the next 5 years in the context of the CLARIAH-PLUS project.

The next 10 years

We will promote and stimulate the use of the applications and services mentioned in the previous sections and try to get them to become an integrated part of the standard linguistics curriculum. We regularly organise courses and training sessions on these applications. In addition, we plan to extend the facilities for accelerating linguistic research and to increase the empirical base of linguistic research. Our plans cover three domains: (1) more and more advanced facilities for text corpus applications; (2) facilities for quickly finding linguistic argumentations; (3) facilities for sharing example sentences while at the same time offering a glossing service.

1. new facilities for searching and analysing research corpora. We will create a system in which frequencies of a wide variety of linguistic properties are generated for any construction that a linguist might be interested in. This will enable the linguist to test hypotheses and theories on this construction against a large amount of data. We also want to make it possible for such a system to automatically search for relations and correlations, especially unexpected relations, which are hidden in the data and that a linguist is unlikely to suspect related. Such phenomena are sometimes discovered using standard linguistic research techniques, but we want to make sure that we discover all such hidden relations, and computational and statistical techniques can assist us in this.
   The corpus applications can not only be used to investigate adult language use but also for research into language acquisition. It can even (I hope) be used as a basis for setting up experimental language acquisition simulations, which, if successful, will surely provide a lot of evidence about what must be known a priori and what can or must be acquired from the input data. It is not obvious that such simulations can be set up in a sensible way, and it is not clear that any such approach will work, but we will have to investigate this, and the corpus applications and data provide some essential ingredients for it.
   Finally, I hope that these corpus applications can also be beneficial for the study of languages for which no technology is available to linguistically enrich their text corpora. It should be investigated whether this can be done using a parallel corpus between two languages, for one of which a treebank is available. If we abstract from the purely language-specific aspects of the treebank,

it might act as a proxy for a semantic representation. With this we hope to be able to apply the language acquisition simulation techniques mentioned above to acquire properties of a language for which we have nothing but a (sufficiently large) parallel corpus containing a text in a different language for which a treebank does exist.

2. Linguists often have to determine whether a particular word or phrase is of a specific syntactic category (e.g. NP or PP), bears a particular grammatical relation (e.g. object or predicate), bears a particular morpho-syntactic feature value (e.g. singular v. plural), or is of a particular semantic category (e.g. mass v. count). Arguments for determining this or for distinguishing different cases exist but are hidden in the linguistic literature. We aim to identify these arguments in the literature (initially manually, later automatically or semi-automatically) and create a database of LInguistic DIAgnostics (LIDIA), with appropriate metadata, so that linguists can quickly find relevant arguments in favour of or against a particular analysis. It is our hope that this will accelerate linguistic research.
3. Linguistic examples and native speaker judgements about them play an important role in linguistic research. However, such linguistic examples are mostly hidden in the texts of linguistic journal articles. This makes it difficult to find them, to access them and to reuse them, i.e., they are not 'FAIR' (Wilkinson et al. 2016). It would be much better if such examples are stored in a publicly available database with appropriate metadata, and suitable search facilities. At the same time, linguists usually have to make a gloss for each example sentence and add a translation. This is not very exciting work and technically often not so simple. With EXCALIBUR[4] we aim to offer a glossing and translation service for example sentences, which assists the linguist in creating glosses and translations for them while at the same time incorporating each of them in a central example sentence database for reuse by other linguists and for further improving the glossing service.

The next 50 years

Though I must admit that I have a crystal ball in my study,[5] looking into it does not help me predict the future. But I am confident that linguistic research will continue to be dominated by two problems that are so big that I sometimes call them *miracles*, i.e. truly happening events for which we have no clue how to

4. EXample sentences CALIBrated for Use in Research
5. https://surfdrive.surf.nl/files/index.php/s/u10UdLIiQoJ1XEZ

account for them. In computational linguistics, there is the *little miracle*: the fact that humans can easily and effortlessly correctly disambiguate a highly ambiguous linguistic expression. In linguistic research, there is the *big miracle*: how very young children acquire their native language easily and effortlessly, in a short period of time and based on relatively little input data, even though each natural language is an enormously complex system, of which professional linguists after a life-long study may only hope to understand just a tiny fraction.

Though it is almost a certainty that in most of the next 50 years linguistics will have to be carried out without the active involvement of Noam Chomsky, I am confident that the influence of his work on linguistics will continue unabatedly.

References

Odijk, Jan & Arjan van Hessen (eds.). 2017. *CLARIN in the Low Countries*. London: Ubiquity Press. Open Access. DOI: https://www.ubiquitypress.com/site/books/10.5334/bbi/. https://doi.org/10.5334/bbi

Wilkinson, Mark D. et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3.160018. https://doi.org/10.1038/sdata.2016.18

Address for correspondence

Jan Odijk
Utrecht Institute of Linguistics
Utrecht University
Trans 10
3512 JK Utrecht
The Netherlands
j.odijk@uu.nl