Tour de CLARIN

CLARIN Knowledge Centre for Treebanking

Edited by Darja Fišer and Jakob Lenardič
Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents the CLARIN Knowledge Centre for Treebanking and is organized in two sections:

- Section One presents the members of the Knowledge Centre and their work
- Section Two includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the Knowledge Centre's infrastructure in their research

CLARIN Knowledge Centre for Treebanking

Introduction .............................................................................................................. 4
Interview | Helge Dyvik ................................................................................................. 6
CLARIN Knowledge Centre for Treebanking

Introduction
Written by Koenraad De Smedt and Jan Hajič

On June 25, 2015, CLARIN recognized a virtual Knowledge Centre for Treebanking operated by a consortium consisting of the following:

1. the CLARINO Bergen Centre at the University of Bergen, Norway;
2. LINDAT/CLARIN at the Charles University in Prague, Czech Republic.

The aim of the CLARIN Knowledge Centre for Treebanking is to provide support for researchers interested in the following:

1. building, depositing, and/or disseminating their treebanks;
2. exploring existing treebanks available at the consortium.

Knowledge is transferred through training events, written documentation, personal advice, hands-on assistance and resource hosting. The members of the consortium offer serviced open platforms for constructing, managing and exploring treebanks.

Access to some treebanks is open, while access to others is restricted to registered users after signing in. Users can be authenticated through single sign-on at members of the CLARIN Service Provider Federation or of eduGAIN.

The CLARINO Bergen Centre operates INESS (Infrastructure for the Exploration of Syntax and Semantics), an integrated treebanking environment with the following online services:

- accessing, searching and visualizing treebank data in various formats (dependency, constituency, LFG, HPSG);
- building LFG treebanks by parsing and discriminant disambiguation;
- editing dependency treebanks.

Most services can be accessed using a web browser, but uploading treebanks and grammars requires manual support. Currently INESS has more than 200 treebanks available in more than 70 languages.

Written knowledge-sharing on the INESS site includes the following:

- a welcome page;
- a page for getting started;
- an overview of the extensive documentation (including walkthrough and documentation of grammar, query language, web interface, annotation and formats);
- an FAQ (list of frequently asked questions);
- a user forum;
- a list of publications;
- links to related information (including a video and slides from a demo);
- project background.

INESS provided interactive knowledge-sharing at the following events:

- Tutorial at the CLARA Thematic Course on Consolidating and Harmonizing Treebank Annotation, Prague, 2010
- INESS Training Workshop, Solstrand, 2013;
- INESS Training Workshop for NAOB, Solstrand 2014;
- INESS Training Workshop at the ParGram meeting, Warsaw, 2015;
- INESS Training Workshop, Solstrand, 2016;
- tutorial on Multiword Expressions in Treebanks at the 2nd PARSEME Training School, La Rochelle, 2016 (with written notes);
- workshop at MONS, Solstrand, 2017.

LINDAT (Prague) offers interactive services to:

- deposit treebanks in a repository;
- visualize treebank data using Treex;
- search and visualize the treebanks using PML-TQ;
- search treebanks using Kontext.

Written knowledge-sharing at LINDAT includes the following:

- a step-by-step guide for depositing resources;
- an FAQ (list of frequently asked questions);
- a user forum.

Until now, the user forums at the Knowledge Centre have been little used, but the organized events mentioned above have been well attended. Knowledge transfer through personal contact with experts at the Knowledge Centre has proved important for projects aiming at curation of their resources.
What is your role in INESS? Could you describe the main goals of this project?

I am one of the developers of INESS, which stands for Infrastructure for the Exploration of Syntax and Semantics. This project, which began in 2010, had two main goals. The first one was to establish an infrastructure for various treebanks across languages. There are now around 400 treebanks, large and small, in INESS, covering about 70 languages. The second was to develop the first Norwegian treebank based on ‘deep’ parsing, which is what most of my work is related to. I was mainly responsible for the further development of the grammar NorGram, in continuous interaction with the annotators (or more properly, the disambiguators) working with the disambiguation of the parse forests of sentences. The treebank, which is called NorGramBank, currently covers around 80 million words (and with that size is obviously for the most part stochastically disambiguated). This is quite a large number for a syntactically parsed corpus, and it is still growing, as we have fashioned it to be a dynamic resource. In 2015, INESS was – in cooperation with the Czech infrastructure LINDAT, which also specializes in the development of treebanks – recognized as a Knowledge Centre in the CLARIN Knowledge Sharing Infrastructure.

What are the main goals of INESS as a CLARIN Knowledge Centre?

We have started actively working on making the search facilities of INESS more user-friendly, which is one of our main goals as a CLARIN Knowledge Centre. Paul Meurer, who was awarded the 2017 Steven Krauwer Award for CLARIN achievements, has developed a querying system called INESS Search. This query language is very powerful and can handle various syntactic frameworks, such as Lexical Functional Grammar, Dependency Grammar and Head-Driven Phrase Structure Grammar.

While the query language itself is in many ways simpler than other existing syntactic query languages, the treebank annotations are so complex that the query task may seem complicated to a novice. To make it more accessible, I have been working on user-oriented, example-based documentation, and with Paul Meurer on developing query templates.

Please describe your academic background.

I have been a Professor of General Linguistics at the University of Bergen since 1983. I studied at the University of Bergen and at the University of Durham, working with Old Norse and Old English phonology and syntax as a graduate student, and also with foundational issues in generative syntactic theory, which became the topic of my PhD thesis. I also did some work in runology, interpreting a number of recently uncovered Medieval runic inscriptions in Bergen. When Lexical Functional Grammar emerged around 1980, I started working within that and some related frameworks. LFG was later used as the annotation framework for the Norwegian and some other treebanks in INESS. I was also involved in some early work in experimental machine translation in Norway. From the late 1990s, I worked on developing an automatic method called Semantic Mirrors, which derives thesaurus-like lexical information from translation corpora. Around the turn of the millennium, Victoria Rosén and I started to develop the first version of the Norwegian Computational Grammar (NorGram), which is a project that we’re still working on and now also involves other researchers, some also from the CLARINO network.

We have started actively working on making the search facilities of INESS more user-friendly, which is one of our main goals as a CLARIN Knowledge Centre. Paul Meurer, who was awarded the 2017 Steven Krauwer Award for CLARIN achievements, has developed a querying system called INESS Search. This query language is very powerful and can handle various syntactic frameworks, such as Lexical Functional Grammar, Dependency Grammar and Head-Driven Phrase Structure Grammar.

While the query language itself is in many ways simpler than other existing syntactic query languages, the treebank annotations are so complex that the query task may seem complicated to a novice. To make it more accessible, I have been working on user-oriented, example-based documentation, and with Paul Meurer on developing query templates.

http://clarino.uib.no/iness/page?page-id=iness-descr
The example-based documentation, which is currently only in Norwegian, is based on the structure of the Norwegian Reference Grammar and examples found there, and shows in a step-by-step fashion how to search for the exemplified constructions. The query templates are ready-made queries with parameters to be supplied by the user, and they are integrated into the search environment itself. You can see examples of such templates, originating in cooperation with lexicographers, if you choose “Select query templates” under the Sketch tab on the INESS webpage (Figure 1). This gives you various query formulas for a wide range of both simple and complex syntactic constructions, which is a useful showcase for grammarians and philologists who are not used to working with more complex query languages. The idea is to be able to aid users by supplying new query templates on demand.

Could you give a simple example of how INESS Search works?

Let’s say that we want to find relative clauses that function as modifiers in nominal phrases, like the bolded clause in the English sentence *The man who is working in the field is my father*. To find such constructions in the treebank, we only need to enter the following search query in INESS search:

\[ \text{NP > CPrel} \]

This instructs the concordancer to look for all syntactic constructions for which the following holds:

1. There exists a tree structure node which is an NP category and there exists a node which is a clause structure (of type CP, a “complementizer phrase”) headed by a relative-clause subordinator, such as *som* in Norwegian.
2. The CP node must be embedded within the NP node, which is specified by the > operator.

Technically, this query is an abbreviation of the quantified expression \( \#x_:\text{NP} > \#y_:\text{CPrel} \), where the operator \( \# \) stands for an existential quantifier that binds a variable for which a specific categorial property is defined (in this case, NP for variable \( x \) and CPrel for variable \( y \)). However, Paul Meurer has simplified the query language so that it is not necessary to refer to the tree nodes with quantified variables if the variables aren’t used more than once in the query.

The results of such a search query are shown in Figure 2:

Figure 1: Query templates in INESS Search

Figure 2: Using INESS search for identifying relative clauses embedded in NPs
Additionally, it is possible to formulate search queries that take into account both the C-structure and the corresponding F-structure information, as in the following example:

\[
\text{NP} \gg \text{#x} \gg \text{#f}, \text{CLAUSE-TYPE} \text{‘rel’}
\]

In this case, the operator \(\gg\) denotes the mapping from a C-structure node \text{#x} to an F-structure \text{#f}, that contains the grammatical information \text{CLAUSE-TYPE} \text{‘rel’}. This means that the concordancer is now looking for all tree nodes that are embedded within \text{NP} and whose F-structure contains the value \text{‘relative’} for clause type. This search query now allows us to find relative clauses in NPs both with and without overt subordinators (e.g. \text{påstanden du nevnte}, “the claim you mentioned”, where the subordinator \text{som} is omitted). Since the latter types of relative clauses lack the CP layer in the LFG representation on account of the omitted subordinator, it is more complicated to search for them by only referring to their C-structure, as in the case of the simpler query \text{NP} \gg \text{CPrel}. 

The first result is the sentence \text{En politibetjent opplyser til Adresseavisen at det er rundt ti turister som har druknet på Karon Beach hittil i år}, which roughly corresponds to English “A police officer informs the Adressa newspaper that there have been around ten tourists who have drowned at Karon Beach so far this year”. So, the clause \text{som har druknet på Karon Beach hittil i år} ("who have drowned at Karon Beach so far this year") is the relative clause embedded in the NP \text{turister} ("tourists"), which is what the query was looking for. Clicking on the example leads you to a tree representation of its C(ons constituent)-structure and a representation of the corresponding F(eature)-structure, which lists the grammatical features of the nodes in the tree and shows their grammatical functions.
Such syntactic constructions would be much more difficult – if not impossible – to extract from a corpus that isn’t syntactically parsed, since you wouldn’t be able to specify any kind of syntactic relations in the query language. As a Knowledge Centre, we also provide help with formulating new search queries, so if a researcher is interested in any kind of syntactic or to some extent semantic phenomenon but doesn’t know how to extract it from the treebanks, he or she need only contact us. Additionally, if researchers are interested in a more detailed explanation of the kinds of formal relations that underlie INESS, we have prepared a short walkthrough in English that explains the basic idea behind the query language. A fuller documentation of the query language is also provided.

Have the Treebanks of INESS been used in any successful project?

Helene Uri, who, aside from being a linguist, is a famous novelist and children’s writer, wrote a book called Hvem sa hva: Kvinner, menn og språk (Who Said What: Women, Men and Language). She discussed the different ways men and women use language, as well as the different ways in which men and women are written about in various types of discourse. Part of her research was done on the Norwegian Treebank, which, in addition to the syntactic dependencies, provides semantic representations such as predicate-argument structure. Specifically, she used the treebank to find out which verbs are mostly associated with female agents and which verbs with male agents. Helene’s book was very successful and she won the Brage Prize 2018 for it.

The NorGramBank treebank has also proven itself important in relation to the rather complex language situation in Norway, as it consists of a wide variety of textual materials like newspaper articles, poplar research and parliamentary debates in both standards. It is for this very reason among the resources used by the Norwegian Language Council, which is responsible for language standardization. In addition, the lexicographic project NAOB in Oslo is using the treebank in the further development of a new comprehensive web-based dictionary of Bokmål which was published last year. The Oslo lexicographers now try to help finance further development of the treebank, since they understand the importance of having an up-to-date resource that can provide relevant examples chosen from the literature based on the actual syntactic use of the dictionary lemmas. There are also other lexicographic projects using NorGramBank.

You have conducted a few linguistic analyses of your own by using the Norwegian treebank. Could you discuss some noteworthy examples?

I am running a blog in which I discuss grammatical phenomena based on the NorGramBank treebank. I focus on some of the well-known syntactic constructions, since there are many misconceptions about their usage in popular discourse. For instance, in Norwegian, one of the stylistic pieces of advice that you hear time and time again is to avoid the passive, the reason being that it supposedly makes a sentence less informative by omitting the agent. However, such advice often isn’t accompanied by any contextual justification, so it boils down to a prescriptive rule that doesn’t hold water if you look at how passives are used in actual texts. In the treebank, I’ve noticed that passives are especially prominent in popular science. In articles from the forskning.no website, the treebank showed that passives were used in almost 25% of the sentences, on average. Looking at their function in relation to the surrounding context, we usually

4http://clarino.uib.no/iness/page?page-id=INESS_Search_Walkthrough
find that their use *in lieu* of the active voice is well motivated. For example, there were many passive sentences like *Disse funnene har ikke vært beskrevet tidligere* (“These findings have not been described previously”), in which the omitted agent of the verbal action is referentially non-specific, which is something that you would expect given that popular science abounds in generalizations. This means that using active variants, like *Researchers have not previously described these findings*, would not make these sentences any more informative. If anything, they would only disrupt the information flow from the perspective of the surrounding discourse. So, when people give out stylistic advice like “avoid the passive”, what they generally overlook is the function of the construction – that is, the passive voice is a device that makes it possible for the writer to adapt the information structure of the sentence to what is prominent in the context; its use is communicatively oriented and in fact very useful in most cases. Besides, Scandinavian languages have an especially rich variety of passive types, something which makes this stylistic advice particularly harmful.

In another blog post, I looked at how sentence complexity (defined as number and degree of embedding of subordinate clauses) varies between different text types. Perhaps the most interesting finding is that the transcriptions of the Norwegian parliamentary debates contain the most complex sentences by far, even more so than the written genres. I was also able to observe an interesting difference between the two Norwegian written standards in the domain of literature. I found that children’s books written in Nynorsk are the least complex, but children’s books in Bokmål contain, on average, more complex sentences than novels in Nynorsk. (However, the limited size of the Nynorsk children’s books corpus is a caveat here.)

**What are the future plans with regard to INESS?**

We plan to expand the corpus of literary Bokmål texts significantly, as part of our cooperation with the NAOB dictionary project, and to continue making the search facilities more accessible, expanding the documentation with ready-made search examples. We also plan on expanding the Norwegian treebank with more texts in Nynorsk.