Exploring Treebanks with INESS Search

Victoria Rosén
University of Bergen
victoria@uib.no

Helge Dyvik
University of Bergen
helge.dyvik@uib.no

Paul Meurer
Uni Research
paul.meurer@uni.no

Koenraad De Smedt
University of Bergen
desmedt@uib.no

Abstract

We demonstrate the current state of INESS, the *Infrastructure for the Exploration of Syntax and Semantics*. INESS is making treebanks more accessible to the R&D community. Recent work includes the hosting of more treebanks, now covering more than fifty languages. Special attention is paid to NorGramBank, a large treebank for Norwegian, and to the inclusion of the Universal Dependency treebanks, all of which are interactively searchable with INESS Search.

1 Introduction

The richly structured information in treebanks requires powerful, user friendly tools for their exploration. We demonstrate the current state of INESS, the *Infrastructure for the Exploration of Syntax and Semantics* (Rosén et al., 2012a; Meurer et al., 2013). The project implementing and operating this infrastructure is carried out by the University of Bergen (Norway) and Uni Computing (a division of Uni Research, also in Bergen), and is funded by the Research Council of Norway and the University of Bergen (2010–2017).

INESS is aimed at providing access to treebanks to the R&D community in language technology and the language sciences. It is is developed and operated in Norway, and has been integrated in CLARINO, the Norwegian part of CLARIN. One of the project's main activities is the implementation and operation of a comprehensive open treebanking environment for building, hosting and exploring treebanks. The other main activity is the development of a large parsebank for Norwegian.

INESS offers comprehensive services for the construction, management and exploration of treebanks. A modern web browser is sufficient as a client platform for accessing, searching and downloading treebanks, and also for the annotation of LFG-based parsebanks, including computer-aided manual disambiguation, text cleanup and handling of unknown words (Rosén et al., 2009; Rosén et al., 2012b; Rosén et al., 2016). These functions are supported by cataloguing, resource management and visualization (Meurer et al., 2016).

INESS has become a valuable service for research groups who have developed or want to develop treebanks, but who cannot or do not want to invest in their own suite of web services for treebanking. Among the larger treebanks developed by others and made available through INESS are the Icelandic Parsed Historical Corpus (IcePaHC, 73,014 sentences) (Wallenberg et al., 2011), the German Tiger treebank (50,472 sentences with dependency annotation, 9,221 with LFG annotation) (Brants et al., 2002) and the dependency part of the Bulgarian BulTreeBank (11,900 sentences) (Simov and Osenova, 2004).

The remainder of this paper provides search examples in recently added treebanks. In Section 2 we present NorGramBank and illustrate search in LFG treebanks. Search in the UD treebanks is illustrated in Section 3.

2 NorGramBank

NorGramBank (Dyvik et al., 2016) is a large treebank constructed by parsing Norwegian text with a wide coverage grammar and lexicon (NorGram) based on the Lexical-Functional Grammar (LFG) formalism (Bresnan, 2001). Approximately 350,000 words of parsed text have been manually disambiguated and checked using computer-generated discriminants. Through stochastic disambiguation the corpus has been extended to about 50 M word tokens. A grammar coverage test was performed on 500 random sentences, of which 78.4% received gold analyses and 6.8% received analyses with only a single local error (Dyvik et al., 2016).

INESS Search is is a querying system for tree-

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 326–329,
Gothenburg, Sweden, 23-24 May 2017.

banks in a variety of formats (Meurer, 2012). It handles search in constituency, dependency and HPSG treebanks as well as in LFG treebanks. The core of INESS Search is a reimplementation of TIGERSearch in Common Lisp and contains numerous extensions and improvements. INESS Search supports almost full first-order predicate logic (existential and universal quantification; negation; quantifier scope can be specified), with the exception of universal quantification over disjunctions (Meurer, 2012, for further details on the query language and the implementation). Partial structures that match variables in queries are highlighted in the interface.

The query language contains some operators that are specific for searching in LFG structures. Among them are the projection operator (to query for c- to f-structure projections), the path operator (to search for f-structure paths satisfying a regular expression over f-structure attributes), the c-command and the extended head operator.

The rich information in NorGramBank allows highly detailed queries. As an example, consider the task — of interest to lexicographers, for example — to study the set of nouns modified by a given adjective, with frequencies. The syntactic expression of such modification may take several forms: attributive position (*a successful result*), simple predicative (*the result wasn't very successful*), object predicative (*they considered the result highly successful*), predicative in a relative clause (*it is difficult to get a result which is completely successful*), etc. Across all these varieties the noun and the adjective always share the value of the feature GEND (gender) by reentrancy in the f-structure representations. This can be exploited in a query like (1), searching for nouns modified in various syntactic ways by *vellykket* 'successful'.

(1)
```
#x_ >PRED 'vellykket' & #x_ >ATYPE &
#x_ >GEND #g_ & #y_ >GEND #g_ & #y_
>(NTYPE NSEM) & #y_ >PRED #p
```

The query says that there exists an f-structure `#x_` which has 'vellykket' as the value of the attribute `PRED` (predicate), has a value for `ATYPE` (i.e., it is an adjective), and has the value `#g_` for `GEND`. Furthermore there exists an f-structure `#y_` which also has `#g_` as the value of `GEND`, has a value for the path `(NTYPE NSEM)` (i.e., it is a noun), and has the value `#p` for `PRED`. The absence of an underscore in the variable name `#p` signals that its values should be listed in the output, which makes it possible to sort the hits by the predicate values of the modified nouns. This gives the output shown in Table 1, showing the top of the list of 355 nouns found, with frequencies. Clicking on one of the lines in the table brings up a display of the sentences found for that word combination.

Count	#p: *value*
24	forsøk
23	prosjekt
18	operasjon
15	resultat
14	behandling
12	kveld
10	1
10	landing
9	menneske
9	eksperiment
8	aksjon

Table 1: Top of the list of nouns modified by *vellykket* 'successful'

Figure 1 shows the analysis of a sentence from the search output with the values of the search variables from the query expression highlighted in red: *Ekspedisjonen ble ansett som vellykket* 'The expedition was considered (as) successful'. The example illustrates how the query expression, based on a shared GEND value, finds examples where the modification relation between adjective and noun is mediated by complex syntactic constructions involving object predicatives, passive, control, etc.

3 The UD treebanks

The Universal Dependencies (UD) project is developing cross-linguistically consistent treebank annotation for many languages.[1] The number of UD treebanks has been increasing dramatically. We have imported and indexed all publicly available UD treebanks (up to v2.0), in order to make them searchable with INESS.

Since all treebanks in this large collection follow the same general annotation principles, they are good targets for showing the capability of INESS Search to search across many treebanks at the same time. For instance, an earlier pilot study (De Smedt et al., 2015) illustrated the use of INESS Search to get a quick indication of the correctness and consistency of annotation across all the UD version 1.1 treebanks.

[1] `http://universaldependencies.org`

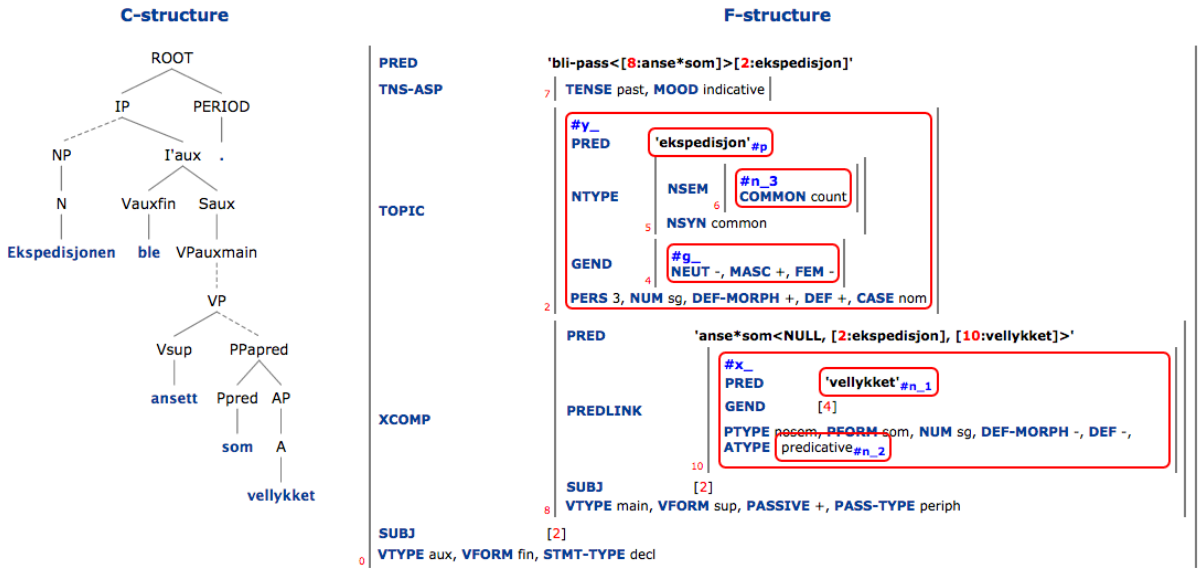

Figure 1: The analysis of a retrieved sentence with the values of the search variables highlighted in red

According to the UD guidelines version 2 on the UD website, the *fixed* dependency relation is one of the three relations for multiword expressions (MWEs). The *fixed* label is to be used for certain fixed grammaticized expressions without any internal structure, but the guidelines do not make it entirely clear whether such expressions must consist of only adjacent words. If one is interested in finding out whether this relation is actually used for annotating non-adjacent words, query (2) can be used to search for binary *fixed* relations that are non-adjacent.

(2)
```
#x >fixed #y & !(#y . #x) & !(#x .
#y) & !(#x >fixed #z & #z != #y) ::
lang
```

Query (2) says that there is a node `#x` which dominates a node `#y` through a *fixed* relation. Furthermore, it is not the case (the exclamation point is the negation operator) that `#y` immediately precedes `#x`, and it is not the case that `#x` immediately precedes `#y`. It is also not the case that `#x` has a *fixed* relation to a node `#z` which is not equal to `#y`.

The result in Table 2 shows the top search results obtained by (2) from the UD v1.3 treebanks for German, Italian, Spanish and Swedish, and also illustrates that the global variable `lang` (for language) can add useful information from the metadata. Figure 2 shows an example of a non-adjacent *fixed* relation in Swedish. The search variables are added in red, making it easy to spot them when inspecting a large dependency structure.

Count	#y: word	#x: word	globals: lang
63	que	de	por
40	ل	ب	ara
23	da	tako	slv
23	sedan	för	swe
22	من	ب	ara
21	quod	quod	lat
19	من	على	ara
19	toe	tot	nld
18	إلى	ب	ara
17	الى	ب	ara
16	menos	a	spa
14	في	ب	ara
13	satunya	satu	ind
12	que	tudo	por
12	ker	zato	slv
11	quod	ita	lat
9	hari	sehari	ind
9	ある	必要	jpn

Table 2: Top search results for nonadjacent *fixed* relations obtained by query (2)

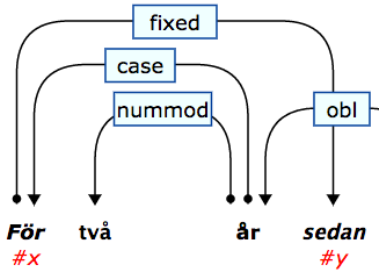

Figure 2: The analysis of a retrieved *fixed* relation with the search variables in red

References

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories*, pages 24–41.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell, Malden, MA.

Koenraad De Smedt, Victoria Rosén, and Paul Meurer. 2015. Studying consistency in UD treebanks with INESS-Search. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 258–267, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Helge Dyvik, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes. 2016. NorGramBank: A 'Deep' Treebank for Norwegian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3555–3562, Portorož, Slovenia. ELRA.

Paul Meurer, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard, and Martha Thunes. 2013. The INESS treebanking infrastructure. In Stephan Oepen, Kristin Hagen, and Janne Bondi Johannessen, editors, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16*, number 85 in Linköping Electronic Conference Proceedings, pages 453–458. Linköping University Electronic Press.

Paul Meurer, Victoria Rosén, and Koenraad De Smedt. 2016. Interactive visualizations in the INESS treebanking infrastructure. In Annette Hautli-Janisz and Verena Lyding, editors, *Proceedings of the LREC'16 workshop VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 1–7. ELRA.

Paul Meurer. 2012. INESS-Search: A search system for LFG (and other) treebanks. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG '12 Conference*, LFG Online Proceedings, pages 404–421, Stanford, CA. CSLI Publications.

Victoria Rosén, Paul Meurer, and Koenraad De Smedt. 2009. LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord, and Koenraad De Smedt, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012a. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Victoria Rosén, Paul Meurer, Gyri Smørdal Losnegaard, Gunn Inger Lyse, Koenraad De Smedt, Martha Thunes, and Helge Dyvik. 2012b. An integrated web-based treebank annotation system. In Iris Hendrickx, Sandra Kübler, and Kiril Simov, editors, *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 157–167, Lisbon, Portugal. Edições Colibri.

Victoria Rosén, Martha Thunes, Petter Haugereid, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer, Gunn Inger Lyse, and Koenraad De Smedt. 2016. The enrichment of lexical resources through incremental parsebanking. *Language Resources and Evaluation*, 50(2):291–319.

Kiril Simov and Petya Osenova. 2004. BulTreeBank Stylebook. BulTreeBank Project Technical Report 5, Bulgarian Academy of Sciences.

Joel Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC) version 0.9.