

LFG PARSEBANKER: A Toolkit for Building and Searching a Treebank as a Parsed Corpus

Victoria Rosén, Paul Meurer and Koenraad De Smedt

University of Bergen and Unifob AKSIS

LaMoRe Research Group

E-mail: {victoria|paul.meurer|desmedt}@uib.no

Abstract

We present the LFG PARSEBANKER, a comprehensive toolkit for interactive incremental construction of a treebank as a parsed corpus. This web-based toolkit offers an environment for batch and interactive parsing, versioning, inspection of structures, discriminant-based disambiguation, and statistics. It has recently been extended with a structural search facility.

1 Overview

In the context of the TREPIL project we have developed methods and tools for the semi-automatic construction of treebanks as parsed corpora. Our approach to treebanking aims for efficiency by combining automatic parsing and computer-assisted manual disambiguation. Parsing is implemented in XLE [5] into which any compatible LFG grammar [1] can be uploaded. Through the automatic identification of discriminants [3], the manual work can remain focused on simple choices even if annotations can be detailed and complex. This method has been successfully applied in earlier research contexts [11, 6], while we have adapted the use of discriminants to LFG structures [9, 7, 8, 10]. We now focus on the toolkit that implements this method.

A practical and efficient implementation of parsebanking with discriminants presupposes an interactive environment that integrates the manual work with automated services such as treebank navigation, computation of discriminants, structural search, versioning, etc. Extending our earlier work [7], we present an updated version of the LFG PARSEBANKER, a comprehensive toolkit that provides this functionality. The toolkit can be used through a web browser and offers the following webpage views:

1. A *treebank overview* page provides corpus selection and navigation in the selected corpus. A partial screenshot from the overview page is shown in figure 1. This view lists all sentences in the corpus together with information such as the sentence ID, the number of parse solutions, the number of discriminants chosen out of the number of total discriminants, the number of chosen analyses, and the sentence length. The annotator may choose to display additional information; among the possibilities are parse time, version information, whether the annotation process is finished, and annotator comments. This view also offers sorting of sentences depending on their various properties.

| <i>Id</i> | <i>Solutions</i> | <i>Disc.</i> | <i>Chosen</i> | <i>Words</i> | <i>Sentence</i> |
|------------------|-------------------------|---------------------|----------------------|---------------------|---|
| 1 | 4 | 1/44 | 1 | 8 | Sofie Amundsen var på vei hjem fra skolen. |
| 2 | 6 | 2/114 | 1 | 9 | Det første stykket hadde hun gått sammen med Jorunn. |
| 3 | 12 | 3/201 | 1 | 5 | De hadde snakket om roboter. |
| 4 | 2+14 | 1/26 | 1 | 11 | Jorunn hadde ment at menneskets hjerne var som en komplisert datamaskin. |
| 5 | 7 | 2/183 | 1 | 10 | Sofie var ikke helt sikker på om hun var enig. |
| 6 | 64 | 6/256 | 1 | 10 | Et menneske måtte da være noe mer enn en maskin? |
| 7 | 16 | 4/178 | 1 | 8 | Ved det store matsenteret hadde de skilt lag. |
| 8 | 30+30 | 3/91 | 1 | 17 | Sofie bodde i enden av en vidstrakt villabebyggelse og hadde nesten dobbelt så lang skolevei som Jorunn. |

Figure 1: LFG PARSEBANKER overview page

2. XLE-WEB is an interface to the XLE parser on a web page.¹ This interface presents the parse results for a single sentence. It includes a display of packed structures and shows the computed discriminants. The user can disambiguate the sentence by selecting or rejecting discriminants and thereby retaining or rejecting sets of corresponding analyses.
3. The *parsebanking* page offers views and disambiguation as in XLE-WEB, but also additional parsebank management operations, such as subcorpus selection. This page also contains a search window (see figure 2).
4. The *discriminant statistics* page presents a frequency list of chosen discriminants for a subcorpus. Each discriminant is listed with its type, the number of times it has been chosen (i.e. marked as good) and the number of times it has been rejected (i.e. marked as bad).

¹<http://decentius.aksis.uib.no/trepil/xle.xml>

Most of these components are implemented in Common Lisp and use XML, XSLT and Javascript to serve the interface web pages. C-structure trees (and graphs) are drawn using Scalable Vector Graphics (SVG) and MySQL is used to store the parsebank, although the system is also compatible with other databases.

2 LFG SEARCHTOOL

The current version of the LFG PARSEBANKER offers a new advanced search facility that allows the user to search for structural characteristics in both c-structures and f-structures. This search tool is modeled after TIGERSearch² [4], which has been reimplemented and extended to cover f-structures, which are not trees, but directed graphs, possibly with structure sharing and cycles. It has a query language that is suitable for c- and f-structures and is more convenient to use than the generic TIGERSearch syntax, to the extent that a graphical interface is not deemed necessary.

In contrast to the TIGERSearch system, whose treebanks and search indices are static, our system allows for dynamic index updating, such that the index always represents the current state of the dynamic treebank. The index data is stored in a database, but in addition, a representation of the index is kept in memory for fast querying.

2.1 The query language

For many common query types it is a problem that they are complicated (or impossible) to express or that they perform unsatisfactorily when using the basic TIGERSearch syntax. Therefore, an abbreviated node description syntax has been devised, and special search language constructs have been implemented that translate to efficient code. Among the new constructs are:

- An abbreviated syntax for c-structure nodes: inner nodes can be specified using the bare node label, whereas leaf nodes are represented by the quoted surface string. For example, query (1) matches all configurations where an NP node dominates a surface node "barna":

(1) NP >* "barna"

- A rule operator '→' to search for c-structure configurations:

(2) NP → N PP

(3) NP → (N #x:PP) & #x → (. P NP .*)

²<http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/>

Query (2) matches all NPs having exactly an N and a PP child, in that order. Query (3) matches the same configuration, where in addition the PP node should have a P node and an NP node as second and third children. The left hand side of the rule operator can be a regular expression over node labels.

- A path operator ‘>(...)’ to search for paths through f-structures (coded as regular expressions over attributes). Examples:

(4) `#f >(COMP TNS-ASP TENSE) "past"`

(5) `#f1 >(TOPIC & OBJ) #f2`

(6) `#f1 >((COMP | XCOMP)+ (OBJ | OBJth)) #f2`

Example (4) searches for a simple path from an f-structure to an atomic value, whereas (5) shows a query involving structure sharing: the f-structure `#f2` should be the value of both the `TOPIC` and the `OBJ` attribute of `#f1`. Functional uncertainty can be expressed using the Kleene star or plus operators, as illustrated in (6).

- A projection operator ‘>>’ relating c-structure nodes and f-structures they project to:

(7) `NP >> #f`

This query finds all NP nodes and the f-structures they project to. The projection operator can be combined with the path operator:

(8) `#c >>(OBJ PRED) "vei"`

2.2 The query interface

Queries can be input both on the overview page and on the sentence page. Long-running queries may be stopped at any time, and a list of links to matching sentences is updated dynamically while a query runs. Matches can be inspected by clicking on a sentence number in the match list. In the c-structure, matching nodes are highlighted, whereas in the f-structure, matching paths are shown.

The screenshot in figure 2 illustrates a view of the parsebanking page with the results of a search. The treebank “sofie” has been searched with the query in (5) above, which specifies that the same f-structure value must fill both the `TOPIC` and the `OBJ` functions, in other words that the sentence has a topicalized object. This treebank has three matches for the query, listed by their sentence ID numbers under the query window. The sentence displayed is “*Ilden*” *ser vi jo*. (“The fire”, we see after all.) The attributes that share the same value are highlighted in the f-structure display.

Treebank: version , annotation-set: , logged in as: victoria (annotator)
 [size: 1143, unparsed: 171, no solutions: 0, fragmented: 347, ambiguity: 89.91 (orig: 509.03), unambiguous: 383 (orig: 63), ambiguous: 607 (orig: 927)]

Grammar: 1.1/Norwegian-newest | Overview | Administration | Discriminant statistics | Documentation

Query: | maximal # of matches:
 #x >(TOPIC & OBJ) #y

3 matches: #577, #672, #927

Sentence #927: "Ilden" ser vi jo.

(1 solutions, 0.08 CPU seconds, 95 subtrees unified; Grammar date: May 06, 2008 11:57; XLE release of Jan 21, 2008 10:36.)

| hide settings |
 Show ambiguous only | Go to next when disambiguated | Go to #: | Don't show structures when more than 20 solutions | packed
 F-structure: Suppress CHECK Show PREDS only | C-structure: Suppress complex categories | Show MRS | Show discriminant weights

Submatches: all, #1

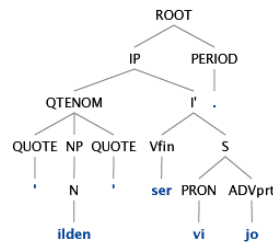
Comment on sentence: -
 On version 1.1: -

Discriminants

Selected solutions: 1 of 1 | gold no good finished

"Ilden" ser vi jo.

C-structure



F-structure

| | |
|---------|--|
| PRED | 'se<[10:vi], [11:ild]>NULL' |
| TNS-ASP | 14 TENSE pres, MOOD indicative |
| PRED | 'ild' |
| TOPIC | 10 NTYPE NSEM 20 COMMON count 10 NSYN common |
| GEND | 17 NEUT -, MASC +, FEM - |
| PERS | 11 PERS 3, NUM sg, DEF +, CASE obl |
| ADJUNCT | 12 { 45 PRED 'jo' } |
| PRED | 'vi' |
| SUBJ | 22 NTYPE NSYN pronoun 10 REF +, PRON-TYPE pers, PRON-FORM vi, PERS 1, NUM pl, DEF +, CASE nom |
| OBJ | [11] |
| | 0 VTYPE main, VFORM fin, STMT-TYPE decl |

Figure 2: Parsebanking page with query result

The toolkit is language independent and can be used with any LFG grammar implemented in XLE, and it has been licensed to several members of the Parallel Grammar Project [2].

References

- [1] Joan Bresnan. *Lexical-Functional Syntax*. Blackwell, Malden, MA, 2001.
- [2] Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The Parallel Grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan, 2002*.

- [3] David Carter. The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, Rhode Island, 1997.
- [4] Wolfgang Lezius. Tigersearch – Ein Suchwerkzeug für Baumbanken. In Stephan Busemann, editor, *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, Saarbrücken, 2002.
- [5] John Maxwell and Ronald M. Kaplan. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–589, 1993.
- [6] Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. LinGO Redwoods, a rich and dynamic treebank for HPSG. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories*, pages 117–128. Växjö University Press, 2003.
- [7] Victoria Rosén, Koenraad De Smedt, Helge Dyvik, and Paul Meurer. TREPIL: Developing methods and tools for multilevel treebank construction. In Montserrat Civit, Sandra Kübler, and Ma. Antònia Martí, editors, *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 161–172, 2005.
- [8] Victoria Rosén, Koenraad De Smedt, and Paul Meurer. Towards a toolkit linking treebanking to grammar development. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 55–66, 2006.
- [9] Victoria Rosén, Paul Meurer, and Koenraad De Smedt. Constructing a parsed corpus with a large LFG grammar. In *Proceedings of LFG’05*, pages 371–387. CSLI Publications, 2005.
- [10] Victoria Rosén, Paul Meurer, and Koenraad De Smedt. Designing and implementing discriminants for LFG grammars. In Tracy Holloway King and Miriam Butt, editors, *The Proceedings of the LFG ’07 Conference*, pages 397–417. CSLI Publications, Stanford, 2007.
- [11] Leonoor van der Beek, Gosse Bouma, Robert Malouf, and Gertjan van Noord. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University, 2002.