

Introduction

We present the LFG Parsebanker, a comprehensive toolkit for interactive incremental construction of a treebank as a parsed corpus. The tool which we have developed supports the process flow in semi-automatic treebank construction, as illustrated in the following scheme:



The toolkit has the following components:

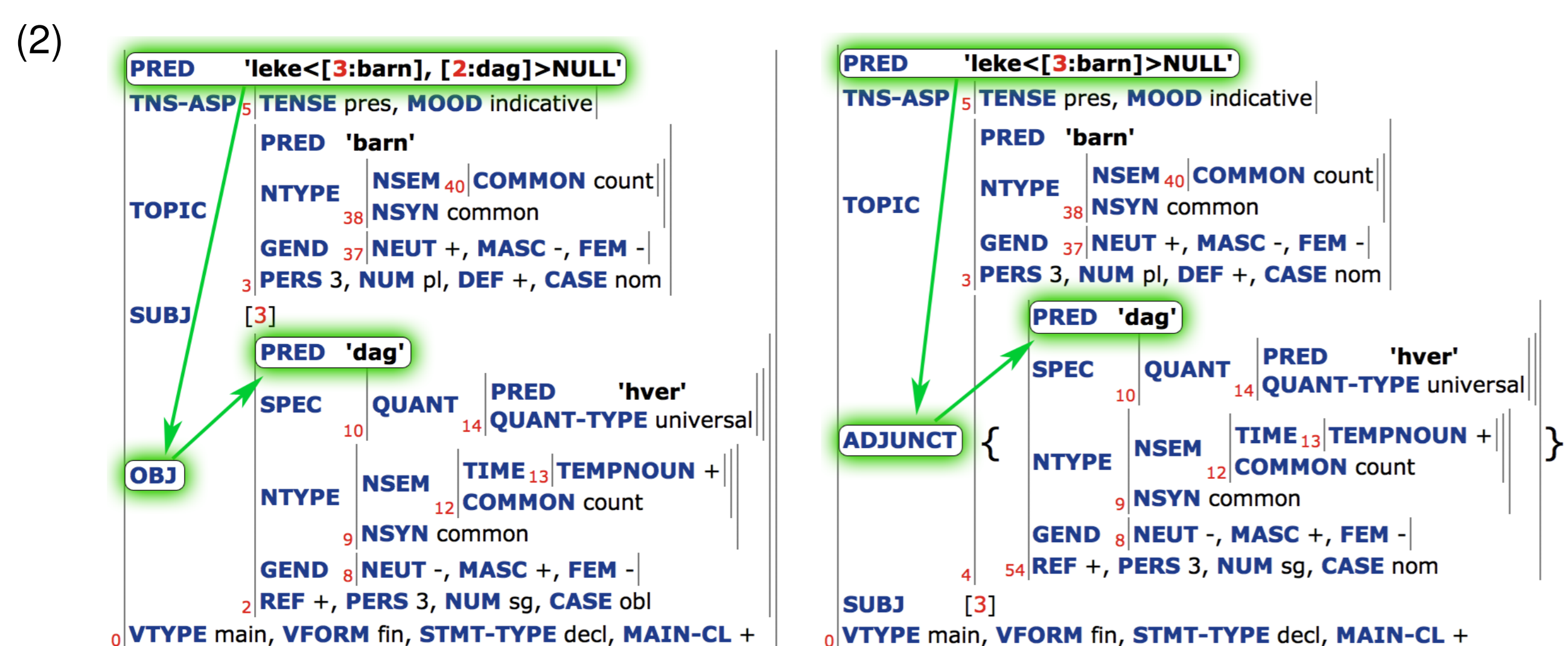
- XLE-Web, an interface to the XLE parser on a web page; this interface includes a new display of packed structures and offers discriminants [1], designed and implemented for LFG grammars, to select an analysis;
- a parsebanking page which offers views and disambiguation as in XLE-Web, but also additional parsebank management operations, such as subcorpus and grammar selection and a search window based on TigerSearch extended for f-structures;
- an overview page providing navigation, information and sorting of utterances;
- a discriminant statistics page displaying statistics on chosen discriminants.

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.

Disambiguation with discriminants

In building a treebank, the annotator's choice between different possible grammatical structures is complicated by several factors. A major challenge is the sheer number of possible structures, which may run into the hundreds or thousands for longer sentences. Another challenge is the high level of detail recorded in the structures, which is desirable in the treebank but can be daunting for the annotator. Consider the f-structures in (2) for the sentence in example (1), where *hver dag* can be an object or an adjunct.

(1) *Barn-a leker hver dag.*
child-DEF.PL play every day
"The children play every day."



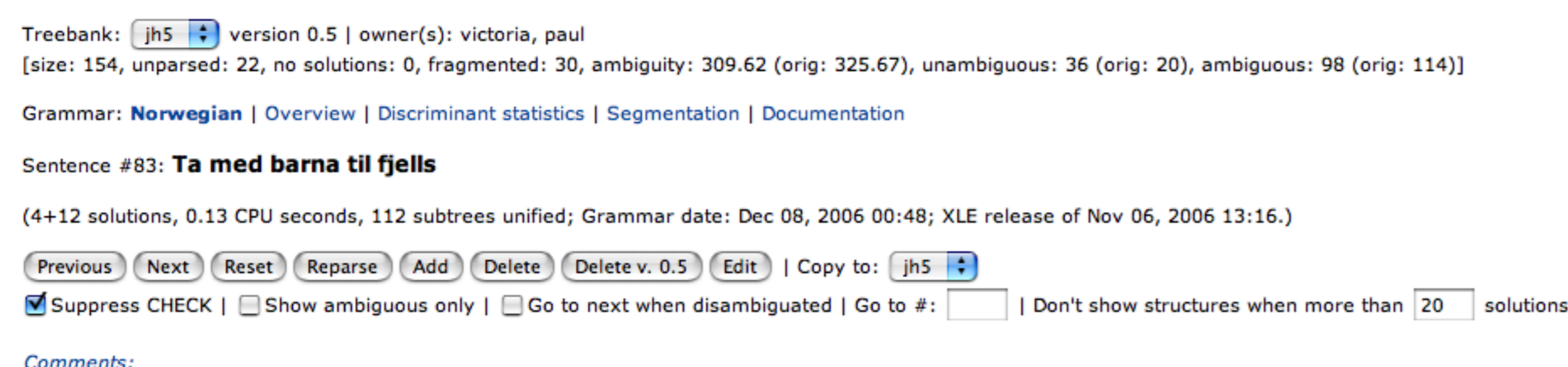
The difference indicated with green shading in the structures in (2) is presented to the annotator as the choice in (3). These simple, local differences are called *discriminants* [1]. By choosing whether *hver dag* is an object or an adjunct, the annotator decides on the intended analysis but avoids examining the whole, complicated structures.



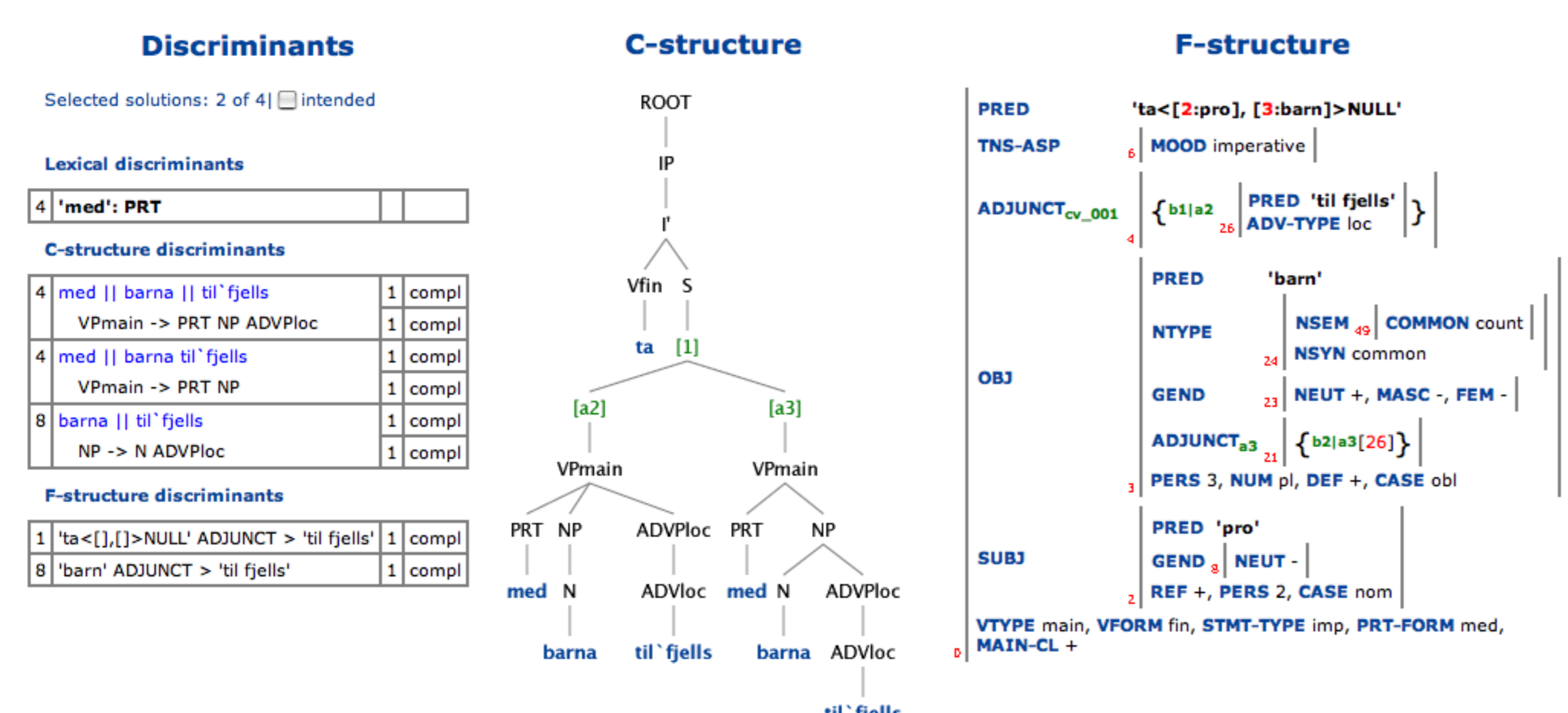
Parsebanking interface with discriminant disambiguation

The interface for identifying the intended analysis is shown in the following screenshot. Here we see the list of discriminants on the left, the packed constituent structure in the middle, and the packed functional structure on the right. The analyses shown are for example (4), in which *til fjells* has two possible attachments. The annotator basically chooses discriminants by clicking to choose or reject them, but other advanced actions are also available [3].

LFG Parsebanker Interface :: Sentence



Treebank: jh5 version 0.5 | owner(s): victoria, paul
[size: 154, unparsed: 22, no solutions: 0, fragmented: 30, ambiguity: 309.62 (orig: 325.67), unambiguous: 36 (orig: 20), ambiguous: 98 (orig: 114)]
Grammar: Norwegian | Overview | Discriminant statistics | Segmentation | Documentation
Sentence #83: **Ta med barna til fjells**
(4+12 solutions, 0.13 CPU seconds, 112 subtrees unified; Grammar date: Dec 08, 2006 00:48; XLE release of Nov 06, 2006 13:16.)
[Previous] [Next] [Reset] [Reparse] [Add] [Delete] [Delete v. 0.5] [Edit] | Copy to: jh5
 Suppress CHECK | Show ambiguous only | Go to next when disambiguated | Go to #: | Don't show structures when more than 20 solutions
Comments:



Discriminants
Selected solutions: 2 of 41 [intended]
Lexical discriminants
4 'med': PRT
C-structure discriminants
4 med || barna || til' fjells 1 compl
VPmain -> PRT NP ADVPl oc 1 compl
4 med || barna til' fjells 1 compl
VPmain -> PRT NP 1 compl
8 barna || til' fjells 1 compl
NP -> N ADVPl oc 1 compl
F-structure discriminants
1 'ta<[]>[]>NULL' ADJUNCT > 'til' fjells' 1 compl
8 'barn' ADJUNCT > 'til' fjells' 1 compl

C-structure
ROOT
IP
f
Vfin S
ta [1]
[a2] [a3]
VPmain VPmain
PRT NP ADVPl oc PRT NP ADVPl oc
med N ADVl oc med N ADVPl oc
barna til' fjells barna ADVl oc til' fjells

F-structure
PRED 'ta<[2:pro], [3:barn]>NULL'
TNS-ASP MOOD imperative
ADJUNCT_{cv_001} {b1:a2} PRED 'til' fjells' ADV-TYPE loc
PRED 'barn'
NTYPE NSEM NSYN common
GEND NEUT +, MASC -, FEM -
ADJUNCT_{a3} {b2:a3[26]}
PERS 3, NUM pl, DEF +, CASE obl
PRED 'pro'
GEND NEUT -
REF +, PERS 2, CASE nom
VTYPE main, VFORM fin, STMT-TYPE imp, PRT-FORM med, MAIN-CL +

(4) *Ta med barn-a til fjells.*
take along child-DEF.PL to mountain-LOC
"Take the children along to the mountains" or "Take the children in the mountains along"

Discriminant types

1. Lexical discriminant (a word form and its part of speech)
2. Morphological discriminant (a base form with its tags from morphological preprocessing)
3. C-structure discriminant (a labeled or unlabeled bracketing of a substring)
4. F-structure discriminant (a minimal path through an f-structure)

Treebank overview page

The overview page, shown in the following screenshot, lists all sentences in the corpus together with information about number of parse solutions, whether the analysis is fragmented, number of discriminants, number of chosen analyses, sentence length, and whether the chosen analysis is the intended one. Any comments added by the annotator during the disambiguation process are also shown.

Treebank: jh2
[size: 319, fragmented: 79, no solutions: 84, ambiguity: 291.91 (280.49), unambiguous: 8 (+49), ambiguous: 227 (-49)]
Grammar: Norwegian bokmål | Discriminant statistics

<i>Id</i>	<i>Sol.</i>	<i>Frag.</i>	<i>Disc.</i>	<i>Chosen</i>	<i>Words</i>	<i>Int.</i>	<i>Sentence</i>	<i>Comments</i>
12	96	*	6 of 127	1	15		Familien Kvame drev hotellet helt fram til 1974, da det ble solgt til Eidsbugarden Turistsenter.	This will get a full parse if '1974' is allowed to take a CPTmprel.
15	8		2 of 132	1	11	*	Det er merkede fotturruter til Gjendebu, Torfinnsbu, Olavsbu, Skogadalsbøen og Yksendalsbu.	
16	8		2 of 43	1	9	*	Vinjestova, forløperen for Eidsbugarden hotell, ble åpnet i 1868.	
17	40		3 of 116	1	14	*	Hotellet ligger i Vang kommune i Oppland, 1060 m o.h., og har 50 senger.	
19	1		0 of 0	1	12		Året etter, da DNT fylte 125 år, ble så turisthytta Fondsbu åpnet.	'Året etter' should be analyzed as NP.
21	40		4 of 114	1	19		De 26 sengene som var i turisthytta ble raskt for få, og det ble nødvendig å bygge et anneks.	CONJspecial should not be allowed before CONJdisc.
31	1		0 of 0	1	13	*	Når du først er i dette området, er også Uranostind et flott turmål.	
32	2		1 of 8	1	5	*	Også den toppen krever brevandring.	
35	30		3 of 131	1	9	*	Det er bilvei til Fondsbu og båtute over Bygdin.	
36	4		1 of 41	1	6	*	Fondsbu turisthytte ble åpnet i 1993.	

Discriminant statistics page

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 is chosen (i.e. marked as good) and the number of times its complement is chosen (i.e. marked as bad). (Note: The statistics shown were compiled before lexical discriminants were added to the system.)

Treebank: jh1
[size: 329, fragmented: 81, no solutions: 71, ambiguity: 377.00 (287.30), unambiguous: 8 (+47), ambiguous: 250 (-47)]
Grammar: Norwegian bokmål | Overview

Discriminant Types:

C(R):	C-structure rule discriminant	(56 [4 as complement])
C(C):	C-structure constituent discriminant	(2)
F:	F-structure discriminant	(232 [57 as complement])
M:	Morphology discriminant	(26 [15 as complement])

Chosen Discriminants (together 316):

Type	Count	Compl	Discriminant
F	21	0	'den' DET-TYPE article
F	13	0	'en' DET-TYPE article
C(R)	13	0	PP -> P YEAR
F	9	9	'ture' NTYPE NSYN common
C(R)	7	0	PROPP -> PROP N PP
C(R)	5	0	IP -> NP I'coord
C(R)	5	0	NP -> N PP
F	5	0	_TOP 'exist<[]>[]'
C(R)	4	0	FRAG -> CONJ FRAG
C(R)	4	0	QuantP -> ART NP
C(R)	4	0	ROOT -> IPimprs PERIOD

Results and prospects

Our work builds on previous parsebanking efforts such as the Treebanker [1], Alpino [4] and LinGO Redwoods [2]. Our toolkit, however, is specifically designed for LFG grammars. We have implemented TIGER-based search on f-structures as well as c-structures, and we can train parse ranking based on our LFG discriminants.

The tool which we have developed is functional and will be further developed in the remainder of the project. Although it was originally primarily intended for Norwegian, it has been implemented in a language-independent fashion. This means that it may be used for building a treebank for any language for which a suitable LFG grammar is available.

The TREPIL project runs from April 1, 2004 to December 31, 2008. Its website is: <http://gandalf.aksis.uib.no/trepil/>.

References

- [1] 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.
- [2] Stephan Open, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. LinGO Redwoods, a rich and dynamic treebank for HPSG. *Research on Language & Computation*, 2(4):575–596, December 2004.
- [3] 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.
- [4] Leonora Van der Beek, Gosse Bouma, Robert Malouf, and Gerjant Van Noord. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University, 2002.