Identifying complex phenomena in a corpus via a treebank lens

Dan Flickinger
University of Oslo
danf@ifi.uio.no

1 Introduction

While syntactically annotated corpora known as treebanks have been available for many years, along with a variety of customized tools for querying these annotations, the mapping from actual annotations to relevant syntactic or semantic phenomena has been obscured by the coarse-grained labelling of nodes in the parse trees which make up the treebanks. This lack of linguistic detail has hampered the use of such treebanks as tools in developing large-scale NLP applications which depend on deep processing, where detailed knowledge about the frequency of occurrence of phenomena within a representative corpus could help in prioritizing domain-specific extensions or corrections to the hand-built grammar resources used in such applications. This paper presents a method for quantifying frequencies of relevant linguistic phenomena within a corpus, by using a Redwoods-style treebank containing rich syntactic and semantic annotations for the corpus, and establishing an explicit mapping between the annotations and the phenomena. Applying this method to the 90,000-word English section of the development corpus used in the LOGON Norwegian-English machine translation project (Lønning et al., 2004) results in a linguistic profile which should highlight a number of development opportunities for the project's deep grammar resources, and thereby improve the end-to-end performance of the demonstrator system.

2 Redwoods Treebank and the ERG

Developed initially within the LinGO (Linguistic Grammars Online) laboratory at Stanford University, the Redwoods Treebank Oepen et al., 2002 is a treebank comprised entirely of analyses derived from a broad-coverage computational grammar, the LinGO English Resource Grammar (ERG; Flickinger, 2000). The ERG is a large-scale HPSG implementation, actively developed at Stanford since 1993, and its analyses provide precise, fine-grained syntactic and semantic information; Minimal Recursion Semantics (MRS; Copestake, Flickinger, Pollard, & Sag, To appear) is the general framework used for meaning representation. Building on an array of existing software tools for processing with the ERG (and similar grammars), the Redwoods Treebank was constructed by parsing selected domain corpora and subsequently hand-inspecting analyses and selecting the intended reading(s) for each input item. Annotation (i.e. manual parse selection) in Redwoods builds on the notion of *elementary discriminants* (Carter, 1997), basic properties of sub-constituents in the parse forest that account for contrasts (i.e. local sources of ambiguity) among analyses. Discriminants—competing lexical entries, for example, or a choice of using the head – complement vs. head – adjunct schema to build a token phrase—are fairly easy to judge, even for non-experts, and enable annotators to navigate the parse forest quickly. Using a specialized tool, each annotator decision on accepting or rejecting a discriminant directly results in the elimination of large parts of the parse forest, so that a small number of local decisions typically will be sufficient to disambiguate even highly ambiguous inputs.

Previous releases of Redwoods included some 15,000 sentences (and sentence fragments) from two domains: transcribed di-

alogues about appointment scheduling and travel planning from the VerbMobil project (Wahlster, 2000); and customer service email messages deriving from commercial development of an automated email response product using the ERG. The most recent release of the treebank, Redwoods 6 (`www.delph-in.net/redwoods`), includes an additional section of 5000 sentences (75,000 words) drawn from the LOGON bilingual development corpus in the Norwegian tourism domain, described in the next section.

As a result of the manual disambiguation of each parsed sentence in the corpus, each of the 20,000 treebanked items in Redwoods 6 includes rich syntactic and semantic annotation, including the detailed derivation tree showing which of the ERG's lexical and syntactic rules were applied to produce the chosen analysis, as well as the full semantic representation in MRS which was compositionally constructed in tandem with the syntactic analysis. Moreover, the rules identified in the derivation tree are instances of construction types represented in the ERG as a multiple inheritance type hierarchy, enabling important generalizations over sets of related grammar rules; and similar abstractions are explicitly represented in the type hierarchies which define the several hundred distinct lexical types which identify the "part-of-speech" for each token in each sentence. Finally, the semantic predicates identifying the elementary predications in the MRS for each sentence also support abstraction through a combination of a type hierarchy for closed-class predicates and a rigorous naming convention for all open-class predicates. It is these fine-grained syntactic and semantic annotations and their underlying type hierarchies which provide the basis for the phenomenon quantification method presented in Section 4. But first, a brief overview of the LOGON corpus used in this study.

3 LOGON development corpus

In order to establish a firm empirical base for its research and development of a linguistically deep machine translation system based on semantic transfer, the LOGON project acquired the rights to a set of booklets on back-country tourism in one of Norway's most popular regions, Jotunheimen. The booklets, originally written in Norwegian, already had one professional English translation, and the project contracted for two additional expert translations, resulting in a Norwegian corpus of about 30,000 words and a corresponding sentence-aligned English corpus of about 90,000 words, with an average length of about 14 words per item (sentence or discourse fragment) in the English corpus. Here are some typical items from one of the English translations:

The start of the hike follows the trail to Gjendesheim through the saddle between eastern and western Hestlægerhø.
Aside from Vestfjorddalen at Rjukan, Gjende, Norway's most beautiful mountain lake, was DNT's principal development area in the early years.
In 1867 a log cabin was put up next to an old stone hut at Nybua, about halfway down the 17.5-mile-long Bygdin, the biggest lake in Jotunheimen.
Owners: Charlotte and Eiliv Sulheim.

To construct the treebank for this corpus, the 6407 items were presented to the ERG[1] for exhaustive parsing, using the PET parser (Callmeier, 2000) and the [incr tsdb()] profiling system (Oepen & Carroll, 2000), also open-source software available from the DELPH-IN website. For the 5738 items which received at least one parse, each was manually treebanked using the discriminant-based approach and tools as usual for Redwoods. Of the 5738 parsed items, 650 had received no correct parse, resulting in a treebank of 5088 items where the one intended analysis is identified and stored, including its derivation tree and its MRS representation. The averages for lexical and syntactic ambiguity assigned by the ERG, along with coverage levels, are shown for one portion of this corpus in Table 1, where the items are grouped by sentence length.

[1]The Jan-06 version was used for this study, since this version is the one employed for all of Redwoods 6. Cf. `www.delph-in.net/erg`

Table 1: Coverage profile for one fourth of the Jotunheimen English corpus, showing the distribution of items by word (token) length, the average lexical and syntactic ambiguity, and grammar coverage.

'gold/erg/jh4' Coverage Profile						
Aggregate	**total items** ♯	**word string** ϕ	**lexical items** ϕ	**distinct analyses** ϕ	**total results** ♯	**overall coverage** %
55 – 60	3	56.67	251.67	23848.00	1	33.3
50 – 54	2	50.00	238.00	0.00	0	0.0
45 – 49	6	46.50	242.17	150758.40	5	83.3
40 – 44	17	42.00	221.73	66813.30	10	58.8
35 – 39	26	36.69	178.56	61130.45	20	76.9
30 – 34	61	31.80	163.43	29185.54	44	72.1
25 – 29	116	26.86	128.11	17409.33	98	84.5
20 – 24	195	21.72	110.25	6131.30	164	84.1
15 – 19	271	17.06	91.20	1005.57	245	90.4
10 – 14	241	11.77	56.03	101.74	234	97.1
5 – 9	295	6.71	31.63	19.06	285	96.6
0 – 4	366	2.16	6.02	2.60	356	97.3
Total	**1599**	**13.60**	**67.05**	**4747.42**	**1462**	**91.4**

(generated by [incr tsdb()] at 14-feb-2006 (15:53 h))

4 Quantifying phenomenon frequencies

Given the Jotunheimen section of the Redwoods treebank, we now have a wealth of annotations to draw from in building a profile of the linguistic phenomena represented in this corpus. The relationship between annotations and phenomena can be a direct mapping from the rule name labelling one node in a tree to a specific phenomenon, as for example with measure-NPs like in *17.5-mile-long*, for which the ERG employs two phenomenon-specific constructions (syntactic rules). But since the ERG is a grammar within the HPSG (Head-driven Phrase Structure Grammar (Pollard & Sag, 1994)) framework, many of the most frequently used syntactic rules are more schematic, with the phenomenon-specific constraints contained in lexical types, such as for subordinate clauses like *... because the route is easily seen*, where the general-purpose Head-Complement, Adjunct-Head, and Subject-Head rules are used to build the clause phrase by phrase. Such lexically-anchored phenomena, accounting for a large number of the ones syntacticians have studied, are also straightforward to identify in the treebank, since the lexical type for each token can be recorded as part of the derivation tree for each item.

A third type of mapping between annotation and phenomenon which is common for these ERG-based trees involves a small set of closely-related constructions which encode minor syntactic variations within the range of one general linguistic phenomenon, such as for relative clauses, which can be finite or not, contain a relative pronoun or not, etc. (see Sag, 1997 for the underlying linguistic analysis). Another readily accessible example of this clustering of constructions involves the handful of distinct rules used in the ERG for building the filler-head phrases which signify the 'top' of an unbounded dependency, with slight but important differences for matrix and non-matrix WH-phrases, topicalized non-WH clauses, and relative clauses. To identify broad phenomena like these within the treebank, the rule names used in the treebank annotations

can be grouped automatically by making reference to the phrasal type hierarchy defined in the ERG itself.

Finally, there are more complex phenomena involving interactions among those which are directly represented by types in the ERG, where identification will necessarily involve testing for one or more configurations within each parse tree. A familiar example involves 'across-the-board' extraction in coordinate structures, as in *a hotel which they modernized and expanded*, where it is the interaction of the (directly encoded) coordination structure with the (directly encoded) NP-gap within each VP which is of interest. For these more linguistically interesting interactions among basic phenomena, involving combinations of constructions and lexical types, it may prove useful to adapt existing treebank-searching tools like those used for more conventional treebanks, as discussed in the following section. However, experience so far suggests that a wide variety of the linguistic phenomena central to the development of applications using deep grammars like the ERG can be identified by establishing relatively direct mappings between a given phenomena and a small set of corresponding annotations.

While the quantification of phenomena and the manual validation/evaluation of this method are ongoing, Table 2 summarizes frequency counts of some familiar linguistic phenomena represented in the LOGON English corpus. Of course, each item in the corpus represents many linguistic phenomena, so the frequency counts and percentages simply indicate the number of items within which the particular phenomenon is present, as identified by the relevant annotation mapping. In addition, these counts are based on the 80% of the corpus which survived parsing and treebanking, and there will clearly be phenomena which are not currently analyzed thoroughly or even at all by the ERG, so the frequency counts do not exhaustively characterize the number of occurrences of even these phenomena in the full corpus. Fortunately, the size of this corpus, while interesting, is tractable enough so that manual inspection of the non-treebanked items can be carried out for some useful set of phenomena, for the practical needs of the LOGON project.

Table 2: Frequencies of occurrence of some familiar linguistic phenomena within the LOGON Jotunheimen English corpus, based on the Redwoods 6 treebank.

Phenomenon Frequency in LOGON		
Phenomenon	♯Items	%Corpus
Measure NPs	279	4.4
Appositives	275	4.4
NP Fragments	1588	24.8
NP Coordination	987	15.4
Multi-NP Coord	265	4.1
VP Coordination	411	6.4
S Coordination	588	9.1
Relative Clauses	486	7.6
Unbounded Deps	1168	18.2
Yes-No Questions	7	0.1
WH Questions	42	0.6
Imperatives	219	3.4
Free relatives	101	1.6
Passives	1072	16.7

5 Related work

Discussion of related work in treebanks and their associated querying tools will be included in the full paper, including discussion of the Penn Treebank and `tgrep/tgrep2` (Pito, 1993), (Rohde, 2001) as well as `CorpusSearch` (Randall, 2000); the International Corpus of English and `ICECUP` (Wallis & Nelson, 2000); the NEGRA corpus and `TIGERSearch` (Lezius, 2002); and the Alpino dependency bank (Bouma, Noord, & Malouf, 2001), as well as recent developments such as the TREPIL project for constructing LFG-based treebanks.

6 Use of linguistic profiles in LOGON

Using the method presented here for determining phenomenon frequencies based on an existing manually-constructed treebank, detailed analysis of a sizeable corpus is readily available. Such information should prove to be useful within the LOGON machine trans-

lation project in at least two respects: (1) in prioritizing those high-frequency phenomena for which special care must be taken in the analysis-transfer-generation pipeline; and (2) in helping to identify potentially problematic phenomena which are frequent in those items where end-to-end translation fails. Both uses will be explored and evaluated as the more detailed linguistic profile of this corpus emerges through refinement and application of the approach described here.

References

Bouma, G., Noord, G. van, & Malouf, R. (2001). Alpino. Wide-coverage computational analysis of Dutch. In W. Daelemans, K. Sima-an, J. Veenstra, & J. Zavrel (Eds.), *Computational linguistics in the Netherlands* (pp. 45 – 59). Amsterdam, The Netherlands: Rodopi.

Callmeier, U. (2000). PET — A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG)*, 99 – 108.

Carter, D. (1997). The TreeBanker. A tool for supervised training of parsed corpora. In *Proceedings of the Workshop on Computational Environments for Grammar Development and Linguistic Engineering.* Madrid, Spain.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (To appear). Minimal Recursion Semantics. An introduction. *Journal of Research in Language and Computation.*

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering, 6 (1)*, 15 – 28.

Lezius, W. (2002). *Ein suchwerkzeug für syntaktisch annotierte textkorpora.* Unpublished doctoral dissertation, University of Stuttgart, Stuttgart, Germany.

Lønning, J. T., Oepen, S., Beermann, D., Hellan, L., Carroll, J., Dyvik, H., Flickinger, D., Johannessen, J. B., Meurer, P., Nordgård, T., Rosén, V., & Velldal, E. (2004). LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation.* Uppsala, Sweden.

Oepen, S., & Carroll, J. (2000). Performance profiling for parser engineering. *Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG)*, 81 – 97.

Oepen, S., Toutanova, K., Shieber, S., Manning, C., Flickinger, D., & Brants, T. (2002). The LinGO Redwoods Treebank. Motivation and preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics.* Taipei, Taiwan.

Pito, R. (1993). *Tgrepdoc man page* (Technical Report). Philadelphia, PA: University of Pennsylvania.

Pollard, C., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar.* Chicago, IL and Stanford, CA: The Univeristy of Chicago Press and CSLI Publications.

Randall, B. (2000). *Corpussearch user's manual* (Technical Report). Philadelphia, PA: University of Pennsylvania.

Rohde, D. (2001). *Tgrep2* (Technical Report). Pittsburgh, PA: Carnegie Mellon University.

Sag, I. A. (1997). English relative clause constructions. *Journal of Linguistics*, *33*(2), 431 – 484.

Wahlster, W. (2000). *Verbmobil: Foundations of speech-to-speech translation.* Springer–Verlag Berlin Heidelberg New York.

Wallis, S., & Nelson, G. (2000). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, *5*(4), 305 – 336.