C-Structures and F-Structures for the British National Corpus

Joachim Wagner, Jennifer Foster, Djamé Seddah and Josef van Genabith
National Centre for Language Technology, Dublin City University
{*jwagner, jfoster, josef*}*@computing.dcu.ie, dseddah@paris4.sorbonne.fr*

We describe how the British National Corpus (BNC) [4], a 100 million word balanced corpus of British English, was parsed into Lexical Functional Grammar (LFG) [8] c-structures and f-structures using a treebank-based parsing architecture. The parsing architecture uses a state-of-the-art statistical parser trained on the Penn Treebank (PTB) [6, 9] to produce c-structures, and an annotation algorithm [5] to enrich the c-structures with corresponding f-structures. We present some issues encountered in applying the parsing architecture on such a large scale. We describe how 1,000 BNC sentences were manually parsed to produce a c-structure gold standard. We present results obtained by evaluating the c-structures produced by the statistical parser against the c-structure gold standard. Finally, we present results obtained by evaluating the f-structures produced by the annotation algorithm against an automatically constructed f-structure gold standard for the 1,000 hand-parsed BNC sentences. To our knowledge, the BNC has not been parsed in its entirety using a deep linguistic grammar formalism. [1] describe how a broad coverage HPSG grammar of English was applied to BNC sentences – the aim of that work was not to parse the BNC but to test the grammar's coverage using a small subsection of the BNC. Our research demonstrates that it is feasible to provide a reasonably accurate LFG analysis of a very large body of sentences in a robust, non-labour-intensive way.

To facilitate parsing with a PTB-trained parser, some reversible transformations were applied to the BNC data, for example, the varcon package was used to translate all words from British English into American English, and punctuation symbols occurring in the BNC but not in the PTB (soft hyphens, neutral quotes) were replaced by their PTB equivalents. Charniak and Johnson's 50-best reranking parser [6] was used to produce the c-structures. A wrapper was implemented which resumes the parsing process if it fails unexpectedly. 99.8% of the 6 million BNC sentences obtained a parse. On 31 2.4 GHz CPUs, parsing the 6,228,111 sentences took 79.5 hours walltime, i.e. roughly 1.425 seconds per sentence. The annotation algorithm was applied to the most probable c-structure, producing an f-structure for all but one sentence.

Gold standard c-structures were constructed by one annotator for 1,000 BNC sentences. The annotator used the PTB bracketing guidelines [2] and the PTB itself as references. The c-structures produced by the parser achieve a precision, recall and f-score of 83.7%, 83.8% and 83.7% respectively, when evaluated against the gold standard c-structures using the labelled Parseval evaluation metric [3]. Results for the eight most frequent c-structure constituents are shown in the lefthand table of Fig. 1. The parser appears to perform quite well on BNC data, considering that it achieves an f-score of 91.3% on Section 23 of the Wall Street Journal section of PTB and an f-score of 85.2% on the Brown corpus [10]. In order to evaluate the f-structures generated by the annotation algorithm, we followed the established procedure of automatically constructing a gold standard set of f-structures by applying the annotation algorithm to the gold standard c-structures. We then evaluated the f-structures produced by applying the annotation algorithm to the c-structures returned by the parser against the gold standard f-structures by computing precision and recall on the f-structures as sets of term descriptions (following [7]). The f-structures received precision, recall and f-scores of 91.1%, 91.4% and 91.2% respectively. A breakdown of the results for the eight most frequent grammatical functions is shown in the righthand table of Fig 1.

Future work will involve manual correction of the automatically constructed f-structure gold standard so that more confidence can be placed in the f-structure evaluation results. It is expected that the LFG analyses of the BNC sentences will be useful in a variety of NLP tasks, e.g subcategorisation frame extraction.

Constituent Type	Precision	Recall	F-Score
NP	86.8	88.4	87.6
VP	81.6	81.8	81.7
S	80.0	81.8	80.9
PP	80.2	82.1	81.1
SBAR	75.8	77.6	76.7
ADVP	80.3	77.4	78.8
ADJP	67.2	69.5	68.3
WHNP	91.9	96.8	94.3

Grammatical Function	Precision	Recall	F-Score
adjunct	83.3	83.6	83.4
num	96.6	97.3	96.9
pers	97.2	97.9	97.5
obj	90.1	90.4	90.2
subj	89.6	87.4	88.5
tense	97.4	96.3	96.8
det	96.5	96.4	96.4
pron_form	98.5	99.0	98.7

Figure 1: Precision, recall and f-scores for most frequent c-structure constituents and f-structure attributes

References

[1] T. Baldwin, E. M. Bender, D. Flickinger, A. Kim, and S. Oepen. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of LREC-04*, 2047–2050, Lisbon, 2004.

[2] A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. Bracketing Guidelines for Treebank II Style, Penn Treebank Project. Technical Report MS-CIS-95-06, University of Pennsylvania, 1995.

[3] E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proceedings of the 1991 DARPA Speech and Natural Language Workshop*, 306–311, 1991.

[4] L. Burnard. User Reference Guide for the British National Corpus. Technical report, Oxford University Computing Services, 2000.

[5] A. Cahill. *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. PhD thesis, Dublin City University, Dublin, 2004.

[6] E. Charniak and M. Johnson. Course-to-fine N-best-parsing and Maxent Discriminative Reranking. In *Proceedings of the 43rd Meeting of the ACL*, 173–180, Barcelona, 2005.

[7] R. Crouch, R. Kaplan, T. H. King, and S. Riezler. A Comparison of Evaluation Metrics for a Broad-coverage Parser. In *Proceedings of the "Beyond Parseval - Towards Improved Evaluation Measures for Parsing Systems" LREC-02 Workshop*, 67–74, 2002.

[8] R. Kaplan and J. Bresnan. Lexical Functional Grammar: a Formal System for Grammatical Representation. In J. Bresnan, *The Mental Representation of Grammatical Relations*, 173–281. MIT Press, 1982.

[9] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[10] D. McClosky, E. Charniak, and M. Johnson. Reranking and Self-training for Parser Adaptation. In *Proceedings of the 21st COLING and 44th Meeting of the ACL*, 337–344, Sydney, 2006.