Introduction to
Frontiers in Corpus Annotation

Adam Meyers
New York University
meyers@cs.nyu.edu

A new annotated corpus can have a pivotal role in the future of computational linguistics. Corpus annotation can define new NLP tasks and set new standards. This may put many of the papers presented at this workshop on the cutting edge of our field.

A standard, however, is a double edged sword. A standard corpus urges users to accept the theory of how to represent things that underlie that corpus. For example, a Penn Treebank theory of grammar is implicit in Penn-Treebank-based parsers. This can be a problem if one rejects some aspects of that theory. Also one may object to a particular system of annotation because some theories generalize to cover new ground (e.g., new languages) better than others. Nevertheless, advantages of accepting a corpus as standard include the following:

- It is straight-forward to compare the performance of the set of systems that produce the same form of output, e.g., Penn Treebank-based parsers can be compared in terms of how well they reproduce the Penn Treebank.
- Alternative systems based on a standard are largely interchangeable. Thus a system that uses one Penn-Treebank-based parser as a component can easily be adapted to use another better performing Penn-Treebank-based parser.
- Standards can be built on. For example, if one accepts the framework of the Penn Treebank, it is easy to move on to representations of "deeper" structure as suggested in three papers in this volume (Miltsakaki et al., 2004; Babko-Malaya et al., 2004; Meyers et al., 2004).

It is my view that these advantages outweigh the disadvantages. I propose that the papers in this volume be viewed with the following question in mind: How can the work covered by this collection of papers be integrated together? Put differently, to what extent are these resources mergeable?

The first six papers describe linguistic annotation in four languages: Spanish (Alcántara and Moreno, 2004), English (Miltsakaki et al., 2004; Babko-Malaya et al., 2004; Meyers et al., 2004), Czech (Sgall et al., 2004) and German(Baumann et al., 2004). The sixth, seventh and eighth papers (Baumann et al., 2004; Çmejrek et al., 2004; Helmreich et al., 2004) explore questions of multilingual annotation of syntax and semantics, beginning to answer the question of how annotation systems can be made compatible across languages. Indeed (Helmreich et al., 2004) explores the question of integration across languages, as well as levels of annotation. (Baumann et al., 2004) also describes how a number of different linguistic levels can be related in annotation (pragmatic and prosodic) among two languages (English and German). The ninth and tenth papers (Langone et al., 2004; Žabokrtský and Lopatková, 2004) are respectively about a corpus related to a lexicon and the reverse: a lexicon related to a corpus. This opens up the wider theme of the intergration of a number of different linguistic resources.

As the natural language community produces more and more linguistic resources, especially corpora, it seems important to step back and look at the larger picture. If these resources can be fit together as part of a larger puzzle, this could produce a sketch of the future of our field.

References

M. Alcántara and A. Moreno. 2004. Syntax to Semantics Transformation: Application to Treebanking. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

O. Babko-Malaya, M. Palmer, N. Xue, A. Joshi, and S. Kulick. 2004. Proposition Bank II: Delving Deeper. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

S. Baumann, C. Brinkmann, S. Hansen-Schirra, G. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich.

2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

M. Çmejrek, J. Cuřín, and J. Havelka. 2004. Prague Czech-English Dependency Treebank: Any Hopes for a Common Annotation Scheme? In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

S. Helmreich, D. Farwell, B. Dorr, N. Habash, L. Levin, T. Mitamura, F. Reeder, K. Miller, E. Hovy, O. Rambow, and A.Siddharthan. 2004. Interlingual annotation of multilingual text corpora. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

H. Langone, B. R. Haskell, and G. A. Miller. 2004. Annotating WordNet. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An Interim Report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

E. Miltsakaki, A. Joshi, R. Prasad, and B. Webber. 2004. Annotating Discourse Connectives and Their Arguments. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

P. Sgall, J. Panevová, and E. Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.

Z. Žabokrtský and M. Lopatková. 2004. Valency Frames of Czech Verbs in VALLEX 1.0. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts.