WHAT TO ANNOTATE?
POSSIBLE FEATURES IN A TREEBANK

April 18, 2007

1 Introduction

This is a list of possible features that one may want to include in a treebank. This list is based on an initial brief list drawn up by Owen Rambow with subsequent additions from Katrin Erk, Chuck Fillmore, Sandra Kübler, Lori Levin, Chris Manning, Joakim Nivre, Martha Palmer, Fei Xia.

Some items have more text associated with them than others. Most of the actual text comes from these subequent additions

This is just a list of things one could include in a treebank, not an endorsement, claim, or demand that these items *should* be included in every treebank.

The plan is to keep adding to this document, ideally with some bibliographic information as well. It might be interesting to list some treebanks and say which features they have. It may be useful to publish this as a tech report at some point.

2 Orthography

- Correcting typos (more important for some languages than others; for example, in Arabic there is a lot of divergence from standard spelling)
- Normalizing regional orthography; for example, in Egyptian Arabic, certain words are consistently spelled differently from the rest of the Arabic world; another prominent example is British and American English
- Segmentation (Chinese, ...)
- Original orthography and/or transliteration into Latin (Hindi/Urdu, Arabic, Chinese, ...)
- Text properties (bold, italic, etc.)
- Text layout, e.g., heading, body text, etc. So far, most processors of treebanks use the sentence as the largest unit, but if we are going to take discourse relations seriously, we need to start processing whole texts, and then text layout becomes important.

3 Morphology

- Super basic POS tag (for languages where some English distinctions are hard to make, say ”nominal” in Arabic which includes Adj, N, and Adv)
- Basic POS tag (N,V,Adj, ...)

- Full morphological tag specifying each component which the words in the language manifest morphologically (number, gender, degrees of comparison, etc)
- Lemma
- Segmentation into subword units with independent syntactic roles. In the Turkish Treebank, words are segmented into so-called inflectional groups (IGs), delimited by derivational boundaries (DBs), each of which has its own part of speech tag. Moreover, syntactic dependencies hold between IGs rather than between words. According to Kemal Oflazer, this is necessary given the incorporating character of Turkish derivation. When a whole sentence consists of one word, incorporating both the subject and the object, a dependency analysis only involving full word forms is not very informative. However, the word forms are of course needed as well, since they impose constraints on possibly syntactic relations. In fact, you have an element of this in English, when you segment *haven't* into *have* and *n't*. It's just that in English this is so marginal that you don't need more than one level.

4 Syntax

- All elements, even if phonetically empty (e.g., empty pronouns)
- Original position of ”moved” elements (i.e., traces)
- Headedness (governor): Just as you distinguish between “moved elements” and “traces” in a traditional phrase structure analysis, you may distinguish between different heads, or governors, in a dependency analysis, e.g., linear head vs. syntactic head (Kahane et al. 1998), or deep governor vs. landing site (Kromann 2006).
- Difference argument/adjunct
- Grammatical function (subject, object, adjunct, ...)
- ”Deep” grammatical function
-
- i.e., regularization of subcat frames for voice, dative shift, etc.
- Headedness (governor)
- Phrase labels
- Difference possession/modification (benefit administration vs benefit's administration)
- All modification relations (even inside NP)
- Difference raising/control
- Difference meaning- bearing preposition vs strongly governed preposition (*I believed in you* vs *In the old days, I believed*) (=CLR dashtag in PTB)

5 Semantics and Pragmatics

- Ellided material. There is some overlap with “All elements, even if phonetically empty” in the syntax section above.
- Predicate- argument structure (à la PropBank/NomBank), without lexicon
- Predicate-argument structure (à la PropBank/NomBank), with lexicon

- Frame semantics (à la FrameNet), without lexicon
- Frame semantics (à la FrameNet), with lexicon
- Lexical meaning in a lexicon (WordNet) and/or in an ontology, and other notions of word sense
- Named entities
- Collocational units with clearly non-compositional semantics; these break down into seveal subtypes:
  - Idioms with little syntactic and lexical variation: *kick the bucket*
  - Idioms, semantically decomposable: *let the cat out of the bag*
  - Metaphoric expressions: *we're at a crossroads*
  - Support verb constructions: *do the dishes ,perform an operation*

  It is very hard to write consistent annotation guidelines about these things, but I would love to see some more annotated data here. I found that an automatic semantic analysis that is blind to these phenomena will stumble quite often.
- Constructions (as in Construction Grammar). Some examples:
  - as smart as her – equative
  - too fast to catch – excessive
  - if I had gone; had I gone – counterfactual
  - What's he doing going to the movies? – incongruity

  These have consequences for natural language understanding in that the meaning is not totally predictable from the structure.

  Possibly in conjunction with a "constructicon" (Fillmore), a compendium of constructions.
- Scope of quantifiers
- Scope of adverbs
- Scope of negation
- Scope of conjunctions and conjunctive adverbs (la Discourse Treebank)
- Co-reference relations. Some of these may fall under syntax if intra-sentential (and be covered by trace annotation, depending on your choice of analysis), but it may also be useful to trace co-reference chains between sentences.
- Discourse relations.

6 Grammars

Treebanks can also be linked directly to grammars, which pre-exist the construction of the treebank, or are constructed in parallel with the treebank.