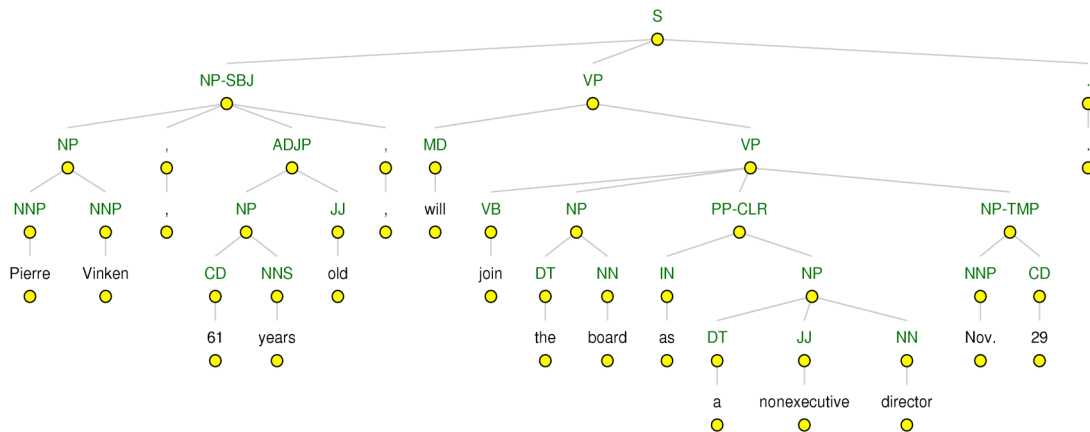


Models of Grammar Evolution: Evolving English

7.1.07

It would be interesting to describe how, for instance, English and French derived from a common ancestor; however, since this is a difficult, novel problem, our goal will be more limited: To develop a model of the evolution of English that can simulate this process. Such an evolutionary model will take a formal model of English and let it change in a plausible way.

This project is a sister-project to “*Models of Grammar Evolution: Biological Applications*”, which had simpler/smaller grammars used in biological applications. Linguistic models are typically much more complex and there is a series of contending approaches. We will single out a particular approach to English and treat that as our universe. The approach is to extract a Probabilistic Context Free Grammar (PCFG) from a “Treebank”: a linguistic resource consisting of sentences of English manually annotated with phrase structure trees. An example phrase structure tree for the sentence *Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.* is given below (taken from slides by Jan Hajic for the Euromatrix project).



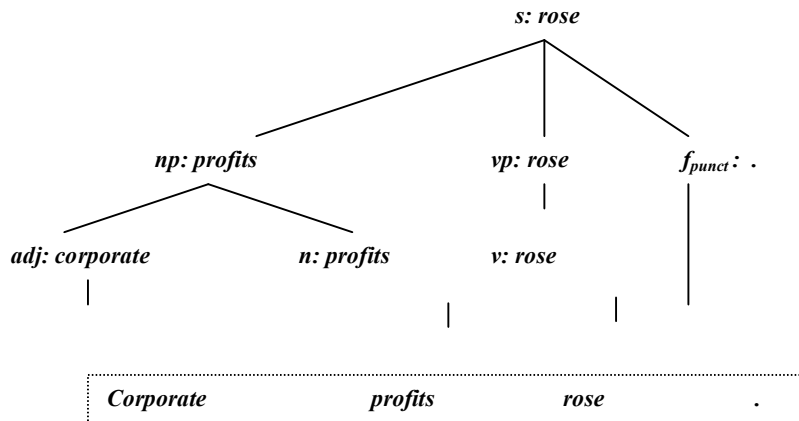
This tree is taken from the Penn Treebank (Marcus et al., 1993), the largest such resource available for English. It consists of 40,000 sentences of newspaper text, each manually annotated with a phrase-structure tree.

Extracting CFG rules from this tree is simple: just take all subtrees consisting of a parent and all its children. For example, the NP subtree on the far left of the diagram gives the following rewrite rule: NP \rightarrow NNP NNP. Deriving conditional probabilities for the rules is also simple: count the number of times that the nonterminal on the left of the rule rewrites as the sequence on the right, and divide by the total number of times the nonterminal on the left appears in the treebank. Charniak (1996) used a similar approach to extract a grammar, which he then used to develop a parser of English. Collins (2003) extended this approach and developed a more accurate parser. A key extension was to include some “lexicalization” in the model, which provides a more accurate model of language. In this case, lexicalization involves passing linguistic “heads” up the tree, so that a non-terminal node also includes the head of the phrase it is governing. For example, adding lexicalization to the tree above would add the word “join” to the VP nodes (since “join” is the head of the verb phrase *will join the board as a nonexecutive director ...*).

The resulting probabilistic context free grammar (PCFG) can then be used to stochastically generate sentences of the language. The number of CFG rules for English obtained by this approach is $\sim 10^4$. The large number of rules is good as it ideally could represent independent characters to be studied. The famous Swadesh (1952) word list that is used in glottochronology is about 200. However, grammar characteristics are more complicated than words and they might not evolve independently.

Synchronous grammars (Wu, 1997 and Chiang, 2005) can analyze two languages simultaneously and are used for automatic translation. In applications to close languages, the two grammars embedded in the synchronous grammar can partially be interpreted as historical transformations of the marginal grammars into each other. Thus the observed events in this transformation can be used for inspiration in defining the basic events that can occur to a PCFG over time. In a synchronous grammar, a rule consists of a pair of CFG rules, one rule for each language, where the symbols are paired across the languages, encoding the translational correspondence. Page 2 of Chiang 2005 contains parse trees for “I open the box” in both English and Japanese (in Japanese the verb comes after the object) together with the resulting synchronous grammar rules.

In modelling the evolution of English we will need to delete/insert rules. Deleting rules is easy, but inserting is considerably more difficult as we have little knowledge of the form a “new” rule will take and so there is much more freedom in how to define insertions.



In this simple example, 3 rules have been applied to generate the sentence (ignoring the generation off words). One could imagine modification by swap for both of these. (ie Corporate profits → Profits corporate). In this project other events could be insertion and deletion. Deleting rules is easy, but inserting is considerably more difficult as we would have little knowledge about how a “new” rule will look and thus we have much more freedom. (example remade from Charniak, 1997)

Project

- Read Charniak (1997), Collins (2003), Marcus (1993) and in Mitkov (eds.) read the chapters on Lexicography (3), Syntax (4), Formal Grammars and Languages (8). Expand the present 2 pages, into 5 page detailed plan.
- Implement a general grammar evolver.
- Measure distance between English and modified English by numerating events happened is one way. These events could not be observed for real languages (only their cumulative effects), but is fine for this idealized project.
- Keep a synch-grammar that summarizes the cumulative effect of the events.
- Find one suitable small text (~400) words) and let it be co-generated in English and modified English. : Compare them for observable differences after 1, 2,4,8,..., 2¹⁵ grammar changing events.

References

- Charniak, E. (1997) “Statistical parsing with a context-free grammar and word statistics”
 Chiang, D (2005) “An Introduction to Synchronous Grammars”
 Collins, M. et al. (2003) “Head-Driven Statistical Models for Natural Language Parsing”
 Knight and Marcu (2004) “Machine Translation in the year 2004”
 Marcus, M. et al.(1993) “Building a Large Annotated Corpus English: The Penn Treebank”
 Mitkov, R. (2005) “Oxford Handbook of Computational Linguistics” OUP
 Swadesh, Morris (1952). Lexico-statistic dating of prehistoric ethnic contacts. Proceedings of the American Philosophical Society 96, 452-463
 Wu, D. (1997) “Stochastic Inversion Transduction Grammars and Bilingual Parsing Parallel Corpora”

Perspectives. This is only a pilot project and there are many interesting issues to pursue:

- Is it possible to calculate a minimal edit distance between two languages?
- The normalized probability of the emitting the same sentences in the two languages?
- How distant will the two languages eventually become in this model?
- It is possible to formulate characteristics/criteria for being a language. Such criteria could restrict which events were allowed in modifying a language grammar.
- The present description is “free evolution”: “English wanders off into language space”, but a step towards real data analysis would be to make directed random walks that forced English to wander towards French. This can be done by MCMC methods (Liu, 2001; Bolhuis et al., 2002) and have been extensively used in biology and elsewhere.
- In real applications, a stochastic formulation should allow investigation if different kinds of rules evolve at different rates.
- It would be interesting to extend the project to higher number of languages (ie English, French and German), this should in analogy with biology introduce the concept of “ancestral language” in a potential triple synch grammar.

Comments. Evolving a language that is described by a PCFG has an analogue in biology, where RNA structures are also described by PCFG and also evolves (Holmes and Rubin, 2002; Holmes, 2004; Dowell and Eddy, 2006)

Additional References

- Bolhuis, PG, -D Chandler, C Dellago and PL Geissler (2002) “TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark” Annual Review of Physical Chemistry Vol. 53: 291-318
 Liu, J. (2001) “Monte Carlo Strategies in Scientific Computing” Springer
 Holmes and Rubin (2002) “Pairwise RNA Structure Comparison with Stochastic Context-Free Grammars” Pacific Symposium Biocomputing 7.163-74
 Holmes, I. (2004) A probabilistic model for the evolution of RNA structure. BMC Bioinformatics.5:166.
 Dowell and Eddy (2006) “Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints” BMC Bioinformatics. 7 (1):400-
 Moulton, V. et al. (2000) “Metrics on RNA Secondary Structures” Journal of Computational Biology 7(1-2): 277-292

Resources

The Collins and Charniak parsers are available from the web pages of Michael Collins (MIT) and Eugene Charniak (Brown). The Penn Treebank is not freely available but we have a copy in Oxford. The Penn Treebank web pages contain some relevant reports and examples.