Comments on Roche

Richard Sproat
Speech Synthesis Research Department
Bell Laboratories, Lucent Technologies
700 Mountain Avenue, Room 2d-451
Murray Hill, NJ, USA, 07974–0636
rws@bell-labs.com

As the original announcement for this Workshop observed, 1996 marks the fortieth anniversary of Chomsky's demonstration of the non-regularity of natural language — more specifically natural language *syntax*.[1] At first glance the computational import of this result may seem obvious: clearly one could not construct a finite-state automaton capable of handling all sentences of a natural language. Yet from the point of view of building practical systems with wide coverage, the import is actually far from obvious. For even if language *were* regular, representing all of the legal sentences of a language via a single finite-state network would be impossible, because the size of the network would be astronomical. As a result, one only finds grammars modeled by single finite-state networks in trivial applications, such as limited-domain small vocabulary speech recognizers, where one can severely limit the set of sentences. Wide-coverage finite state syntactic analyzers must necessarily adopt other strategies.

Apart from Roche's contribution, two other approaches to wide-coverage finite-state syntactic analysis come to mind. One is the Finnish finite-state Intersection Grammar work (e.g., [3]); the other is work on Local Grammars (e.g. [1]).[2]Despite superficial differences, all these approaches are related in that they each involve the intersection of a *text automaton*, containing the possible lexical analyses of an input sentence, with a series of finite-state machines which encode syntactic constraints. This method effectively allows one to sidestep the question of whether language as a whole is finite-state; instead, one performs syntactic analysis with a lot of little constraints, each one of which *is* finite-state. The parallel between such computational approaches to syntax, and theoretical approaches under the general "principles and parameters" rubric is striking: in both cases complex systems of rewrite rules have been abandoned in favor of a scheme whereunder there is a massive overgeneration of output structures, which are then filtered by a series of simple, often lexical, constraints. Roche's approach differs from Intersection Grammars and Local Grammars in one important respect: since he uses FSTs, rather than FSAs, he is able to build structure, as well as restrict it. The primary function of Intersection Grammars and Local Grammars is to weed out possible analyses, and while such analyses may include some bracketings around syntactic constituents (as in the work reported in [3]), the amount of structure 'built' is minimal.

There are a number of questions that one might ask about Roche's particular contribution. Since the parser is defined as T_{dic}^{∞}, one may wonder about the efficiency of performing possibly many compositions of a large dictionary, such as the one Roche reports on in his final section, on a possibly highly ambiguous text automaton. And one wonders why Roche feels that context-free grammars have as a drawback "the inability or the difficulty of handling various types of deletion": methods for handling deletion within strictly context-free formalisms (such as GPSG) have, after all, been known for a long time. But let us sidestep such matters and concentrate on the issue that seems to me to be most interesting in Roche's approach, namely its 'lexicality'. Although the parser is (correctly) described as top-down, it is in one important sense also bottom-up since the table that the parser uses is completely, or almost completely, lexicalized. In other words, each path through the lexicon relates a syntactic frame to one or more lexical items, in a way reminiscent of lexicalized TAGS [2]. As with LTAGs, it is relatively straightforward to think of encoding likelihoods of the various lexicalized syntactic entries, by extending the idea of parsing with transducers to parsing with *weighted* transducers. This in turn suggests a possible alternative approach to the issue of frozen expressions. It is far from obvious to me that one necessarily wants to actually *remove* the 'compositional' analysis for frozen expressions, as Roche proposes. To take a familiar example, one might consider that *kick the bucket*, in addition to its thanatological interpretation, also retains its compositional one. One could handle this by assigning costs that relate to reasonable estimates of the probability of the construction. The cost for the (idiomatic) *kick the bucket* path in the dictionary would correspond to the measured probability of that construction. In contrast, the cost for the non-idiomatic reading might be composed from the unigram cost assigned to the phrase *the bucket*, plus the cost of the lexical frame *NP kick NP*, plus some backoff cost: one could of course further refine these estimates by real ngram estimates, assuming one has data on the likelihood of the non-idiomatic reading of *kick the bucket*. Under this scenario, one would be left with multiple weighted analyses, where the idiomatic reading would surely have the best score, but other readings would be present, and might

[1] This commentary has benefited from discussions with Mehryar Mohri.

[2] Of course, all of these approaches, and especially Roche's, owe some debt to the early work of Woods on ATNs [4]

Proceedings of the ECAI 96 Workshop
Extended Finite State Models of Language
Edited by A. Kornai.

even be selected if other considerations (e.g., pragmatic ones) could decide that a non-idiomatic reading is more appropriate. It remains to be seen how efficiently this could be done within Roche's framework, but a nice property of an FST-based scheme such as Roche proposes is that something of this nature is easy to try.

REFERENCES

[1] Mehryar Mohri, 'Syntactic analysis by local grammars automata: an efficient algorithm', in *Papers in Computational Lexicography: COMPLEX '94*, pp. 179–191, Budapest, (1994). Research Institute for Linguistics, Hungarian Academy of Sciences.

[2] Yves Schabes and Aravind Joshi, 'Parsing with lexicalized tree adjoining grammar', in *Current Issues in Parsing Technologies*, ed., Masaru Tomita, Kluwer, Dordrecht, (1991).

[3] Atro Voutilainen, 'Designing a parsing grammar', Technical Report 22, University of Helsinki, (1994).

[4] William Woods, 'Transition network grammars for natural language analysis', *CACM*, **13**(10), 591–606, (1970).