Mehryar Mohri & Richard Sproat

On a Common Fallacy in Computational Linguistics

Abstract

The question of the computational complexity of natural language has attracted many linguists, formal language theory experts, mathematicians, and computer scientists over the last few decades. To determine the complexity of language, researchers have pointed out various complex morphological or syntactic constructions. But, the arguments used to infer from these observations a particular complexity, in particular the position of a language in Chomsky's hierarchy, are often fallacious. Although this point has been noted many times, such arguments continue to flourish in the literature and even in some computational linguistics courses. The arguments share the same property and depend upon the same common fallacy. While in some cases it is possible to provide a correct formal proof of the complexity of a language, in many cases, the constructions in question cannot be used to demonstrate anything about the complexity of that language. Nonetheless, even in such cases, there is a core intuition underlying the arguments that may be interesting. We remind readers of the common fallacy and sketch how such constructions should be presented and exploited in a rigorous way.

1. The fallacy

Over the last forty years, starting with (Chomsky 1957), there have been a large number of papers addressing the issue of the computational complexity of natural language. Invariably, these take the following form: a particular construction of a language L is discovered that generates a language $L' \subseteq L$ which is at some well-defined position P' in the Chomsky hierarchy, and from this it is concluded that L is at a position $P \geq P'$. Say for example L' is context-sensitive, so L is at least context-sensitive.

The original argument of Chomsky (1957) involved center-embedded constructions, which, for unbounded sentences, are context-free and non-

A Man of Measure
Festschrift in Honour of Fred Karlsson, pp. 432–439

regular.[1] So, from the fact that the language L' of center-embedded constructions is context-free and non-regular, it was concluded that L, English in that case, is at least context-free, which led to the well-known statement (Chomsky 1957: 21):[2]

> English is not a finite state language.

Later arguments in the literature have focused on context-sensitive constructions, such as Cross-Serial Dependencies in Dutch (Bresnan & al. 1982) or Swiss German (Shieber 1985), phrasal reduplication in Bambara (Culy 1985), and more generally morphological reduplication in many languages; see most recently (Walther 1999, Beesley & Karttunen 2000). There has been a general trend in the literature to look for constructions whose properties seem to place them ever higher in the Chomsky hierarchy.

As we shall discuss further below, some of the complexity proofs that have been presented are correct. On the other hand, Chomsky's original argument, and many other statements are, in the form they are stated in, fallacious. The fact that such arguments continue to flourish in the computational linguistics literature suggests that the reason they are fallacious may not be generally understood. Let us cite for example the following excerpt from the abstract of a talk by Hobbs & al. (1997):

> Finite-state models are clearly not adequate for full natural language processing... Every computational linguistics graduate student knows, from the first textbook that introduces the Chomsky hierarchy, that English has constructs, such as center embedding, that cannot be described by any finite-state grammar.

or the following excerpt for instance from a course (Arnold 2000):

> Natural Languages are not Finite State ('regular'). There is no FSA (hence type 3 grammar) that can generate $a^n b^n$. Natural Languages are infinite, and have constructions like $a^n b^n$, i.e. 'nested dependencies'...

[1] We leave aside here the question of the linguistic well-foundedness of such observations.

[2] Note that the correct terminology to use in formal language theory is *regular language* and not *finite state language*. Large classes of context-free languages such as some of those used in Chomsky's argument have been proved to be representable by weighted finite automata (Cortes & Mohri 2000).

Here, again, is why such arguments are fallacious: if a language L contains a subset L' which is at position P' in the Chomsky hierarchy, in general this tells us nothing about the position P of L in the hierarchy. A regular language may well contain a subset that is strictly context-free or context-sensitive. Two of the examples of (strictly) context-free languages given by (Chomsky 1957: 21) are:[3]

(1) $L_1 = \{a^n b^n : n \in \mathbb{N}\}$

(2) $L_2 = \{ww^R : w \in \Sigma^*\}$

But L may contain L_1 or L_2 and yet be a regular language. For example, if $L = \{a^p b^q : p, q \in \mathbb{N}\}$ then L is a regular language that can be recognized by a two-state finite automaton and $L_1 \subseteq L$. More trivially, any language L' based on the alphabet Σ is a subset of $L = \Sigma^*$. Thus, in general, the discovery of a sublanguage $L' \subseteq L$ no matter how complex does not provide any clue about the overall complexity of the natural language L.

2. The correct statement

Naturally it is possible to state the case correctly. One such instance is the treatment of Shieber (1985). Shieber's thesis was that natural languages—or at least Swiss German—show evidence of being weakly context-sensitive, hence not context-free. Swiss German contains a cross-serial dependency construction that has the form:

$v_1 noun_1^m noun_2^n v_2 verb_1^m verb_2^n v_3$,

where $noun_{\{1,2\}}$ are nouns, $verb_{\{1,2\}}$ corresponding verbs and v_i are intervening material. Shieber presented a homomorphism that maps $\mathrm{v}_{\{1,2,3\}}$ to w, x, and y respectively, $noun_{\{1,2\}}$ to a and b respectively and $verb_{\{1,2\}}$ to c and d respectively. Intersecting this with the regular language $wa^*b^*xc^*d^*y$ yields the weakly context-sensitive language $wa^m b^n xc^m d^n y$.

[3] Σ denotes the alphabet, or the set of words of the natural language L to be described. w^R denotes the mirror image of a sequence w. Chomsky (1957) refers to what would later be published as (Chomsky 1959) for a more *formal proof* that English is not a finite state language, but that reference contains no more basis for a mathematical proof of the statement and contains the same fallacious argument.

Context-free languages are closed under homomorphisms and under intersection with regular languages, hence the original language—Swiss German—cannot have been (strictly) context-free. Shieber's contribution is notable in that unlike a previous argument for non-context-freeness of similar constructions in Dutch (Bresnan & al. 1982), the Swiss German facts do not depend upon assumptions about the syntactic structure of the construction. Furthermore, the data seem as linguistically unquestionable as any such data can be, which cannot be said of some other attempts that Shieber cites.

However, while the goal of proving the overall complexity of a given language (or language in general) may be an interesting one, as we saw in Section 1 the mere presence of a construction of a given complexity (as Shieber also clearly notes) says nothing about the complexity of the containing language. Nevertheless, the construction itself may still have an interesting implication for the complexity of a computational device that must recognize instances of that construction, irrespective of the complexity of the language as a whole.

To see this, take for instance the case of full-word reduplication in Malay cited by Beesley and Karttunen (2000). The language L' of full-word reduplicants is strictly context-sensitive (assuming unbounded copy length, on which see Section 3):

$$L' = \{ww : w \in \Sigma^*\}$$

where w is a string over Σ, the set of phonemes of Malay. From the argument in Section 1, we can see that, contrary to Beesley and Karttunen's (2000) implication, this tells us nothing about the complexity of L, the language of well-formed words of Malay.

Note that the infinite language L' is assumed here to be a subset of L. Otherwise, the argument about the complexity of L would be even more fallacious since an observation about the language L' would then be used to draw some conclusion about the complexity of $L \cap L'$. If L is a finite language, for example, the intersection $L \cap L'$ is also a finite, hence regular, language no matter what the position of L' is in the Chomsky hierarchy.

Problems aside, though, there is a point here. Reduplicated words in Malay are not just a subset of the well-formed words of the language. Rather, they are a *well-defined* subset with morphosyntactic attributes that one would like to be able to identify. The full-word reduplication that Beesley and Karttunen (2000) consider, for instance, is a marker of noun pluralization: so we have *orang* ‘person’ and *orang-orang* ‘people’. This suggests that the real problem is not simply to identify or recognize reduplicants as being in the language *L*, but rather to identify and *tag* them distinctively. The output of such a tagger would have all strings in L' tagged with a distinctive tag, say $\boldsymbol{t}$. Let τ be the transduction corresponding to a morphosyntactic tagger of Malay. Assuming for the sake of argument that only reduplicants correspond to plural nouns and thus should be tagged as such, then:

$$\forall\, x \in L',\ \tau\ (x) = x\boldsymbol{t}$$

and $\tau^{-1}(\Sigma^{*}\boldsymbol{t}) = L'$. This implies that τ is not a rational transduction (Berstel 1979, Eilenberg 1974) and cannot be realized by a finite-state transducer since $\Sigma^{*}\boldsymbol{t}$ is a regular language but L' is not. This is because when τ is rational, τ^{-1} is also rational, and rational transductions preserve regular languages.

Two points are critical in this argument: we have not tried to make any implicit claim about the complexity of *L* as a whole; in contrast, what is critical is that the sublanguage L' is a linguistically motivated subset that one would like to tag distinctively (Radzinski 1991).

We have no doubt that this statement of the argument is more or less what many authors have intended in the past when they have made arguments about the complexity of a particular construction in a natural language. However, many presentations of such arguments that we are aware of fail to present the issue in this way.

To summarize, the choice for someone who wants to make a claim about a particular construction’s complexity is clear:

(1) A careful proof of the kind given by Shieber 1985 could be developed, and the data thus used to show the complexity of the containing language as a whole; or

(2) One could back off from this claim and concentrate merely on the construction involved, demonstrating instead the needed complexity of a computational device that must recognize such constructions.

In either case though, one needs to be clear about what precisely is being claimed. Incorrectly presented “demonstrations” merely sow confusion.

3. Further issues

None of the previous arguments is interesting if the length of natural language sentences (or words) is considered to be bounded by some large number *N*. This point is in fact mentioned by Chomsky (1957) in his discussion of the complexity of English. Of course, the size of a finite automaton describing that language may then be large. Interestingly though, other careful syntactic observations about natural languages may suggest that in fact, at a very fine level of description, natural languages should be described by less powerful devices such as finite automata (Gross 1993, Karlsson & al. (eds.) 1995).

The trend of looking for more and more complex constructions to prove some theorem about the complexity of natural language seems to be part of a more general activity that consists of creating new formal theories of natural languages rather than trying to collect more linguistic observations and data. This trend radically contrasts with the work done by scientists in other areas. Did Linnaeus choose to prove a general theorem about the complexity of the development of plants before carefully examining the various features of plants and giving a general classification later used by many to build theories? Computational biologists also give more importance to the collection of large amounts of data about the genome rather than speculating first about a new and general theory. Of course, such epistemological arguments are not needed in the presence of illogical and fallacious arguments.

Acknowledgments

The authors wish to thank reviewers for helpful comments on an earlier version of this paper. Mehryar Mohri’s work was partially funded by the New York State Office of Science Technology and Academic Research (NYSTAR).

References

Arnold, Doug (2000) *LG511 Computational Linguistics I: Parsing and Generation*. University of Essex. URL: http://courses.essex.ac.uk/lg/LG511/1-Formal/index 7.html.

Beesley, Kenneth & Lauri Karttunen (2000) Finite-state non-concatenative morphotactics. In Jason Eisner, Lauri Karttunen & Alain Thériault (eds.) *Finite-State Phonology: Proceedings of the 5th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pp. 1–12.

Berstel, Jean (1979) *Transductions and Context-Free Languages*. Stuttgart: Teubner Studienbucher.

Bresnan, Joan, Ronald Kaplan, Stanley Peters & Annie Zaenen (1982) Cross-serial dependencies in Dutch. *Linguistic Inquiry* 13: 613–635.

Chomsky, Noam (1957) *Syntactic Structures*. The Hague: Mouton.

—— (1959) On certain formal properties of grammars. *Information and Control* 2: 137–167.

Cortes, Corinna & Mehryar Mohri (2000) Context-Free Recognition with Weighted Automata. *Grammars* 3: 2–3.

Culy, Christopher (1985) The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* 8: 345–351.

Eilenberg, Samuel (1974) *Automata, Languages and Machines*. Orlando, FL: Academic Press.

Gross, Maurice (1994) The Lexicon-Grammar of a Language. Application to French. In Ronald E. Asher (ed.) *The Encyclopedia of Language and Linguistics*, pp. 2195–2205. Oxford & New York, NY: Pergamon Press.

Hobbs, Jerry R., Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel & Mabry Tyson (1997) FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In Emmanuel Roche & Yves Schabes (eds.) *Finite-State Language Processing*, pp. 383–406. Language, Speech, and Communication. Cambridge, MA: The MIT Press.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä & Arto Anttila (eds.) (1995) *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing 4. Berlin & New York, NY: Mouton de Gruyter.

Karttunen, Lauri (2001) Applications of Finite-State Transducers in Natural Language Processing. In Sheng Yu & Andrei Paun (eds.) *Implementation and Application of Automata, 5th International Conference, CIAA 2000, London, Ontario, Canada, July 24-25, 2000, Revised Papers*, pp. 34–46. Heidelberg: Springer.

Manaster Ramer, Alexis (1988) Review of Savitch, Walter J.; Bach, Emmon; Marsh, 8, 333–343. William; and Safran-Naveh, Gila (eds.) The Formal Complexity of Natural Language. Studies in Linguistics and Philosophy 33. Dordrecht: D. Reidel, 1987. *Computational Linguistics* 14: 98–103.

Radzinski, Daniel. 1991. Chinese number-names, tree adjoining languages, and mild context-sensitivity. *Computational Linguistics* 17: 277–299.

Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8: 333–343.

Walther, Marcus. 1999. *Deklarative Prosodische Morphologie*. Linguistische Arbeiten 399. Tübingen: Niemeyer.

Contact information:

Mehryar Mohri
Courant Institute of Mathematical Sciences
251 Mercer Street, New York, NY 10012, USA
mohri(at)cs(dot)nyu(dot)edu
http://www.cs.nyu.edu/~mohri

Richard Sproat
Department of Linguistics and the Beckman Institute
University of Illinois at Urbana-Champaign
707 South Mathews Avenue, Urbana, IL 61801, USA
rws(at)uiuc(dot)edu
http://www.linguistics.uiuc.edu/rws