

Directions in Corpus Linguistics

Proceedings of Nobel Symposium 82
Stockholm, 4–8 August 1991

Edited by
Jan Svartvik

Mouton de Gruyter
Berlin · New York 1992

Comments

by *Fred Karlsson*

The general strategy suggested for design of corpus analysis software in the 1990s is laudable. However, the principles of preferring fully automatic analysis over human intervention, and of having processing speed take precedence over ultimate precision of analysis, are potentially risky, at least if applied without due care. They could lead to lowered quality standards.

Even a 1% error rate seems too high in regard to most linguistic phenomena, including the basic task of part-of-speech tagging. This would yield 10,000 errors in 1 million words, 1 million errors in 100 million words. Casual users might be unaware of the limitations of recall and precision of such analysis systems.

An interesting kind of remedy is the use of partial or underspecified analyses. In situations of uncertainty, the analyzer would make no decisions but rather leave the alternatives pending. If the uncertainty rate can be brought, for example, to the 1% level, it is quite possible, for many purposes (such as tagging a corpus), to have these decisions made by human intervention. Only the uncertainties would be presented for human evaluation, for example on-line. Such intervention is fast and dependable, compared to the task of having to spot 1% errors by scanning all of an analyzed corpus.

It is important both for theoretical and practical reasons to have perfect analysis as one of the central goals of automatic corpus analysis.

It is also important to have future natural language processing (NLP) systems more extensively documented, tested, evaluated, and compared to other similar systems, than is presently the case. Another central goal is to promote free or cheap non-commercial scientific use of existing NLP systems. The formalisms should be language-independent (applicable as such to any natural language).

Some brands of theoretical linguistics have not paid enough attention, or only a misguided type of attention, to corpora. But of course there are also instances of corpus-based studies where the problems addressed and the answers given are more or less trivial. "Good" theory and "educated" corpus study should be united throughout the research process. Real problems should be insightfully described on a sufficient level of abstraction.

Mod
parsers
module
made i
exampl
form th
near co
on bou
presupp
parsing
restrict
How
alternat
the wh
the var
that, e
prefixe

The
< CLE
directio
ifier), c
pronou
reading
exampl
one, an
a transi

Usin
on mor
pus stu
has bee
Anttila
alternat

Modularity is a central principle of software design. The idea of "partial parsers", each covering some central subproblem, is a promising one. In NLP, module interconnections and interplay are important. Often, decisions to be made in one module depend upon answers provided by other modules. For example, is there a clause boundary in front of a certain instance of the word-form *that*? A dependable answer presupposes a fairly conclusive parse of the near context, to be performed by another module, which in turn might rely on boundary information. Successful resolution of such interdependencies presupposes, *inter alia*, monotonic accumulation of information during the parsing process, and powerful declarative means of expressing the relevant restrictions.

How should partial parses and indeterminacy be represented in NLP? One alternative is Constraint Grammar Parsing (CGP, Karlsson 1990) viewing the whole enterprise of morphosyntactic parsing as disambiguation. Consider the various morphological readings that could be attributed to the word-form *that*, each on its own line below, with its potential syntactic codes (each prefixed by "@" within parentheses):

```

that <CLB> CS (@CS)
that ADV (@AD-A>)
that <NONMOD> PRON DEM SG (@SUBJ @OBJ @PCOMPL-S
@PCOMPL-O @I-OBJ @<P... )
that DET CENTRAL DEM SG (@DN>)
that <NONMOD> <CLB> <REL> PRON SG/PL (@SUBJ @OBJ
@PCOMPL-S @PCOMPL-O @I-OBJ @<P... )

```

The readings are complementizer (CS = subordinating conjunction, < CLB >= starts a new clause), adjectival intensifier (">" indicating the direction of the head), pronoun in head function (< NONMOD >= not modifier), determiner as modifier of a noun to the right (@DN>), and relative pronoun. The constraints of Constraint Grammar disambiguate morphological readings and syntactic codes equally by way of discarding alternatives. For example, for *that*, the morphological reading "PRON SG/PL" is the proper one, and its syntactic function is @OBJ, in sentence-final position if there is a transitive verb to the left in the same clause.

Using some 1500 constraints of this down-to-earth type, relying decisively on morphological information provided by the lexicon and on extensive corpus studies, a full-scale morphosyntactic parser for English called ENGCG has been developed at the University of Helsinki (Karlsson - Voutilainen - Anttila - Heikkilä 1991). All ambiguities "are there" at the outset, undecided alternatives will be left in the output, constraints discard most (optimally: all

spurious) alternatives. When applied to fresh running text, the error rate of part-of-speech assignment of ENGCG is less than 0.3%.

The CG formalism is a compromise between pure qualitative grammar statements and pure probabilistic descriptions. It is fully language-independent.

Tagging and parsing are closely related. Part-of-speech tagging could be seen as the basic step of any parser. In parser design, lexicon and grammar should be closely integrated. A central module of a robust parser is a Master Lexicon covering the core vocabulary, 30,000-50,000 lexical items. The Master Lexicon should work in conjunction with a proper and precise morphological analyzer yielding inflectional descriptions such as (1), including base-form reduction (lemmatizing), that serve as input to the disambiguation modules of the parser. In this sense, morphological analysis is indispensable in automatic analysis of any language.

One more task facing corpus analysts is determining the proper or optimal corpus size for various types of linguistic problems. The magnitude of this problem grows in parallel with corpus size. For example, if 500,000 hits are retrieved from a 200 million word corpus, powerful software for scanning and structuring the hits will be in great demand.

References

Karlsson, Fred

1990 "Constraint grammar as a framework for parsing running text", in: H. Karlgren (ed.), *Proceedings from the XIIIth Int. Conf. on Computational Linguistics* 3: 168-173. Helsinki.

Karlsson, Fred - Aro Voutilainen - Arto Anttila - Juha Heikkilä

1991 "Constraint grammar: a language-independent system for parsing unrestricted text, with an application to English", in: *Proceedings from the AAAI-91 Workshop on Natural Language Text Retrieval*. Anaheim, CA.

The od
The lin
high q

Henry K

Compute
pact on v
bibliogra
puterized
affected r
word-rec
where the
to more e

My pu
linguistic
and signi
of these
word pro
of them a
respectab
of such l
in the En
our comm
even som
least sma
search fu

As W.
of langua
mation w
word frec
linguistic
corpora f
able form
databases