COMPUTATIONAL LINGUISTICS IN 1990

Hans Karlgren

KVAL Research Institute for Information Science
Skeppsbron 26
S-111 30 Stockholm, Sweden
coling@com.qz.se

Coling's existence today can be measured in (reasonable fractions of) centuries. After twelve well-renowned international conferences arranged by the linguistic behaviour as a manner to better understand how we speak and listen, write and read, learn and unlearn, understand, store and retrieve information. An ultimate question is to what extent these our most human activities can be reduced to mechanistic operations: by teaching machines what we can recognize what in us is machine-like. Whenever we can mechanize something which seems deeply human, we gather urgent, often painful, knowledge about ourselves. Whenever we fail, we may learn even more. It is not only in thermodynamics that the great failures mark the great advances. Cf. the coloquium on The Unfinished Language.

Some colleagues would say computational modelling of human linguistic behaviour is the goal. I think it is going too far to require that computational models of human language must needs be valid as possible [future components of] models of the human intellect; that is a most point of a rather remote philosophical nature since we can as yet rarely ever verify claims about the similarity or analogy between the working of our models and human minds.

One thing I see as crucial in computational linguistics at this particular point of time is machine learning; cf. my portion of the Summing-Up-and-Look-Ahead session at Coling 88 in Budapest, subsequently published along with the other statements of that session in the Prague Bulletin No. 51, which was intended as a seed for COLING-90.

Modelling learning is interesting in itself but modeling language user's learning and adaptation also attacks one of the most salient features of natural languages and one which so far is not well-studied behavior of invented languages: the intriguing feature that human users understand utterances and texts by means of knowledge about the language system and that such knowledge is successively acquired from the utterances and texts we understand.

To get a relevant model for human linguistic competence we must teach machines to learn: to update their grammar and lexicon from the very texts on which they apply them, treating the texts as operands for the analyzers and simultaneously as operators that modify the analyzers. It is my belief that there are basic procedures, as yet poorly understood, which are common to language change over longer periods, language acquisition by an individual and the mutual adaptation between dialogue participants or the reader's adaptation to the author during and possibly merely for the purpose of the current dialogue or text.

The most important successful attempts to handle very large text corpora and huge lexical data bases might obscure this crucial issue and postpone its solution: I feel uneasy about some impressive analyses and syntheses based on sub-sub-subcategorizations of words and situations in some microslice of our world. Close-up on some instances are indispensable in guiding empirical research, but continued fact collecting and algorithm building does not necessarily bring us generalizable insights or generalizable procedures. The conclusion when we have succeeded in mapping some detail, which turned out to be more complex than we could imagine, should not always be to find resources, ours or somebody else's, for every other detail to be mapped with equal precision; but to model the procedure for such mapping.

Details must be seen in a context and I believe that the most fruitful context at the very present is that of learning and adaptation.

Artificial intelligence does study machine learning. But I expect that it is from linguistics, with its tradition of studying change and with an object which so obviously does not wait till the next authorized release before it changes, that a major break-through will come for linguistic adaptation and for learning at large.

None of what I have now said should be taken to mean that applied computational linguistics is unworthy of discussion at Coling. Applications can help us ask new questions, and the successes and, even more, the failures in practical tasks give us very valuable feedback, confirming and disconfirming our beliefs. But it should be clearly understood that practical application is not the ultimate test of the value of what we are doing: I think it is absurd to see, say, the needs for office automation as a justification for our study of human language.

Thus, if somebody would have put together an automatic translator, actually producing readable output when given arbitrary economic or technical prose, the world would not have become a very different place, although quite a few organizations would have run more smoothly; the insights gathered from trying to translate mechanically by mere dictionary and syntax would provide us with essential knowledge of translation, of language and hence of ourselves. In the case of machine translation, therefore, I prefer papers which illuminate some feature of the task of translating which they claim to be (un)programmable to those which demonstrate how well their tool works.

An international conference can be seen as a stimulus-response sequence. The initiators of Coling emit a stimulus to a wide community of people who probe human language - and such as do not know they do - and get a response we can only partially control: we set things in motion by
announcing the conference, we aim at an intended
target area by filtering the contributions of and we can to some little extent guide the missiles
underway by giving directions and hints to the
authors and of others.
How do we judge the result now when the
contributions have arrived? It is obviously
premature to answer the question how Coling 90 will
have amended our concept of computational linguist-
ics or to evaluate the papers, since much of
their content is still in the dawning and premature
corrections. We hope the readers will disagree on a
number of points added in these volumes. In any
case, there is a certain incubation time for really
new ideas to have an effect. But some first-
impression comments from the only one so far who
has read – though in several cases certainly not
yet digested – all the papers, those published here
and those which have been rejected, could make the collection appear less amorphous to some
reader, whether or not he sympathizes with the
views proferred. See, however, my attempt at a
more demanding task than foreseen. Unlike what has
been accepted generally we have rejected the rest, since many of them are
purely quantitative reasons we could accommodate
the most common critical comment by referees on any
category of papers this time which report on the rediscovery of
phonology, text planning, idioms, anacelutha or the
written instruction manual. On the whole, however,
the advances beyond full stop continue, though
perhaps more less obviously than one might have hoped. Most attempts concern cohesion between neighbouring
sentences – such as anaphoric phenomena – but
interesting results have also been achieved in
more apparent form for larger structures (text
planning). – Stylistic studies are rare as are
other studies of literary theory. An
attempts to identify text topics and one on concept
analysis and terminology address issues directly
relevant for document retrieval and linguistic analysis.
4.3 New territories have been invaded in a more
immediate sense. We have the pleasure of seeing
contributions from geographical and linguistic
regions from where we earlier had no reports. Thus, we see an encouragingly large number of papers from the
Chinese-speaking parts of the world.
This expansion also means that our linguistic
models and accepted ideas have been put to a test
on more languages than before. The predominance of
English examples in linguistic research world-wide introduces a bias, the amount of which we cannot,
by nature, possibly estimate but which should worry
us.
4.4 Computational linguistics is moving out of the
laboratories. Today’s computational facilities –
fast on-line operation, large data-bases, con-
venient user-interface prices acceptable even to
front-line research institutes – have made experi-
ments much more life-like. Performance and elegance as reported in the Project Notes are often impres-
sive.
The ability to handle realistic vocabularies has led to a revived interest in lexicology. In
particular, the knowledge accumulated in dictionary
phonological issues is often impress-
sive. As reported in the Project Notes are often impres-
sive.
4.5 Linguists have lifted their gaze to see beyond
the written text. We catch a glimpse of the
interest in speech processing – both
analysis and synthesis – is increasing. Some
reports on ambitious investments in this field was
left out of Coling 90 because the linguistic tools
applied – or shall we say as yet applicable? – were
rather unsophisticated. More interdisciplinary
efforts within computational linguistics may prove
fruitful in a near future. – It may be relevant
here to observe that while there are a few, not
many, contributions on phonology, there is still
none on phonetics.
4.6 Some ventures beyond the map are reported:
Attempts to make an automatic system cope reason-
ably with phenomena which are in some sense
unexpected, especially that the system was designed to
effect unexpected things to happen.

4.2 The advances beyond full stop continue, though
perhaps more less obviously than one might have hoped. Most attempts concern cohesion between neighbouring
sentences – such as anaphoric phenomena – but
interesting results have also been achieved in
4.7 Linguists have lifted their gaze to see beyond
the written text. We catch a glimpse of the
interest in speech processing – both
analysis and synthesis – is increasing. Some
reports on ambitious investments in this field was
left out of Coling 90 because the linguistic tools
applied – or shall we say as yet applicable? – were
rather unsophisticated. More interdisciplinary
efforts within computational linguistics may prove
fruitful in a near future. – It may be relevant
here to observe that while there are a few, not
many, contributions on phonology, there is still
none on phonetics.

4.6 Some ventures beyond the map are reported:
Attempts to make an automatic system cope reason-
ably with phenomena which are in some sense
unexpected, especially that the system was designed to
effect unexpected things to happen.
It is in the nature of things that some of
these attempts are as yet more speculative; cf. the
colloquium on "The Unfinished Language".

4.2 The advances beyond full stop continue, though
perhaps more less obviously than one might have hoped. Most attempts concern cohesion between neighbouring
sentences – such as anaphoric phenomena – but
interesting results have also been achieved in
4.7 Linguists have lifted their gaze to see beyond
the written text. We catch a glimpse of the
interest in speech processing – both
analysis and synthesis – is increasing. Some
reports on ambitious investments in this field was
left out of Coling 90 because the linguistic tools
applied – or shall we say as yet applicable? – were
rather unsophisticated. More interdisciplinary
efforts within computational linguistics may prove
fruitful in a near future. – It may be relevant
here to observe that while there are a few, not
many, contributions on phonology, there is still
none on phonetics.

4.6 Some ventures beyond the map are reported:
Attempts to make an automatic system cope reason-
ably with phenomena which are in some sense
unexpected, especially that the system was designed to
effect unexpected things to happen.
It is in the nature of things that some of
these attempts are as yet more speculative; cf. the
colloquium on "The Unfinished Language".
Automatic acquisition of linguistic knowledge is a topic of several papers, particularly for extension of the lexicon, and also other learning procedures such as automatic derivation of rules from corpora and the successive accumulation of knowledge, in a "knowledge base", for translation.

A very gratifying extension in the analysis of faulty input. Or let us say non-canonical input since it is clear that real persons do not comply exactly with any canon. Not only are 'robust' systems necessary to make computational experiments more realistic (and man-machine interaction in practical systems more human) but the underlying distrust in incommutability in human behaviour is also theoretical implications. Very serious implications, possibly: if there is more to it than can be handled by a little normalization filter, the doubtful old distinction between competence and performance must be revised.

4.7 Concurrently with the renewed interest in discovery procedures statistical procedures have become frequent. Rarely have we seen so many numbers in a collection of Coling papers as this time. An egocentric: reminding: as one who spied a good deal of statistical apparatus to promote statistical methods in linguistics (and publishing a scientific journal, NML, for the purpose) connected but not released at this conference, even though there is no indication of any causal relationship between those efforts and the present trend.

Unfortunately, the methodological level of the majority of experimental and statistical work in linguistics today is no higher than in statistical linguistics 30 years ago. Little efforts are spent on the non-trivial process of constructing appropriate stochastic models for linguistics. And, quite unnecessarily, the accumulated knowledge of the statistical profession is disregarded and "(relative) frequency" are often used indiscriminately as a synonym, a discussion of estimates and sample in most correct and the reader is supplied with uninterpretable percentages.

This methodological weakness is probably due not only to an understandable inacquaintance with the trade. There is also a cultural barrier: Computational linguistics has torn down an important section of the wall between mathematics and traditional humanities but has inherited another, that between mathematics and statistics. Though theoretical statistics is a respected part of mathematics, many top-level mathematicians and logicians remain uninterested in statistical models for linguistics. And, quite unnecessarily, the accumulated knowledge of the statistical profession is disregarded and "(relative) frequency" is often used indiscriminately as a synonym, a discussion of estimates and sample in most correct and the reader is supplied with uninterpretable percentages.

5. To characterize computational linguistics of 1996 we now remarks; necessary, otherwise incomplete and subjective, about what is lacking.

5.1 Computer simulations, as well known in, say, modern physics, are rare. Of course, the whole of our field may be seen as a kind of simulation. A very healthy trend is to make experiments more realistic - with more than a toy dictionary and a restricted grammar and with some amount of extra-linguistic knowledge. But the opposite trend is absent: to simplify down to the barest minimum, i.e., not only to use restricted grammars but to intentionally make false assumptions, as when a physicist describes a gas or a dozen highly idealized realistic experiments are more readily done than a few years ago, but the idealized experiments have not become simpler.

I shall not elaborate this point here but I believe we are still burdened by the humanistic delight and craving for a full-fledged knowledge - and lack of the mathematician's delight in fruitful simplification. We have more barriers to break down than this case that keeps some of us from working with intentional falsification.

5.2 In spite of the interest in discovery procedures, computer-supported factfinding seems rare. Thus, for more and more schemes are proposed on how to make machines do things which need to be done but very little interest has been shown in systematically what need to be done, for instance by recording in some more than superficial manner what humans do.

5.3 Language change has been left out of discussion, except for comments in An unfinished language. We have seen nothing about historical linguistics: not even on diachronical phonology which seemed promising some years ago. This would have a still other reason than the inherited institutional structure in the scholarly community, and I expect great progress when top competence in historical linguistics is combined with the insights and tools of computational linguistics.

5.4 For similar reasons, probably, we have seen nothing in the philological field. The challenging field of manuscript reconstruction, for which we would by now have received no computational linguist this time, nor has authorship attribution. We have already remarked on the emptiness of the borderland between linguistic and literary studies.

5.5 No contribution, as we already noted above, took up a phonetical topic. This gap is so much we were again inclined to view research was heavily computerized at an early stage.

6. Finally a deeply felt appeal to all Coling participants and all more-than-casual readers of this publication: Do not forget that we are in the humanistic domain which could be more humane than examining the frontier between what in us is man and what in us is machine when we perform our most human activity.

Many papers contain a trace of bad conscience: we know we should do more for our applicable work; we ought to justify what we cost now and if society only allows us time for basic research we shall certainly deliver.

This apologetic attitude is groundless: unjustified and unjustifiable. Your allegiance is not primarily to investors. True, if you have accepted funding to perform practical tasks, you must try to live up to your promises. But most of the resources dedicated to our field do not come from those who represent future users or uses.

In particular this is true for the resources which have gone into Coling. The Coling conferences are arranged by the International Committee for Computational Linguistics which never received a penny from anybody and owes obedience to no political or commercial body, national or international. The support provided to each conference by governments and industry is welcome but represents a minute portion of the total effort in real terms, as do the participation fees.

To take one detail of which I have first-hand knowledge, the monies spent by the referees, at least three of which examined and commented upon each paper submitted, which in turn outnumbered those now published by a factor of ten. That is, there is a million-dollar affair if evaluated at market price for comparable consultations (an evaluation which admittedly requires some stretch of imagination); not one referee even asked whether he or she would be paid or get a reduced participation fee. All resources for the preparation, work done by the international committee, our Finnish hosts and first of all by the writers, and where my own six manmouths, unpaid but highly rewarding, is a small trifly by comparison), absence (from production?) at home, travel and accommodation costs and all secondary and tertiary expenditure, all these make a conference of this kind a multi-million venture which could readily have rescued villagefuls of starving people from a painful death, bought a privileged country an extra submarine to fight its enemies - were supplied to us for a cause. They were not intended to benefit the industry or to the administrations, which can no doubt afford to pay for the development they need.

So: Do not trivialize your pursuit! You have more serious business to do than business.