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 can today be measured in (reasonable fractions of) centuries. After twelve well-renowned international conferences arranged by the International Committee for Computational Linguistics and after an increasing number of publications and local meetings dedicated to the topic one might assume that all of us who care would by now know exactly what computational linguistics is about. We do not. Not only do we differ slightly on where we want to place the emphasis, the concept also evolves with each of us in the vague, successive way in which other words in human language so fruitfully keep changing. Computational linguistics is what we make it.

No doubt computational linguistics is about computation and linguistics, with an emphasis on 'and'. The key concepts are computation, not computer, and linguistics, not language processing. All agree that the scope does not extend to studies, however good, about computation applied to language material unless some linguistic insight is at issue or about computer support for linguistics unless the computational procedure has some non-trivial linguistic aspect.

What, then, is the core of the matter? I can give only a personal answer; let me do that. Thereafter I shall comment upon the harvest of papers offered to Coling 90.

A great goal is to model computationally human linguistic behaviour as a manner to better understand how we speak and listen, write and read, learn and unlearn, understand, store and restructure information. An ultimate question is to what extent these our most human activities can be reduced to mechanistic operations: by teaching machines 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 colloquium 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 moot 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 "processing".

One theme which 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 conspicuously absent from 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 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-ups on some instances are indispensible in serious 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 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 offered and we can to some little extent guide the missiles underway by giving directions and hints to the authors, speakers and discussants.

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 linguistics or even to evaluate the papers, since much of their value lies in the debate and protests they provoke. We hope the readers will disagree on a number of points argued 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 could not be accommodated, 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 selective subject index.

1. The first observation about this year's Coling papers is that there are many of them. Computational linguistics is no more an interdisciplinary oddity. It has grown into a discipline.

2. The papers are on a whole on a high level. For purely quantitative reasons we could accommodate less than one in three. We have avoided to say that we have rejected the rest, since many of them are valid and interesting contributions to the field and would have been accepted for Coling had there been more time and space in our conference week.

The quality was disturbingly good from the immediate viewpoint of the referees, who have had a more demanding task than foreseen. Unlike what has happened at earlier Coling conferences, very few contributions could be dismissed because they were trivial computerizations of linguistic studies or software engineering achievements with amateurish linguistic assumptions. We have seen very few papers this time which report on the rediscovery of the fact that nouns forms in some language can conveniently be categorized in number, gender and case or which specify file structure and computational environment while leaving the linguistic issues undiscussed. Most computational linguists today know what they are talking about and take appropriate professional hardware and software tools for granted.

3. Like other established domains, computational linguistics risks to become isolationistic. By far the most common critical comment by referees on any category of papers was that the authour had overseen or left unmentioned relevant earlier work. It is a symptom of what Bertrand Russel called Provincialism in Time if for a topic of a general nature the list of references includes only items from the last decade. Such a narrow perspective may be natural if one discusses technicalities in a recently presented formalism or some implementation details, but not if one investigates, say, dialogue, text planning, idioms, anacolutha or the lexicological treatment of hapax legomena or non-canonical forms.

4. The papers submitted to Coling 90 report on significant advances on many frontiers.

4.1 Much work has been done on the refinement and elaboration of the conceptual apparatus. I do not yet venture to point at innovation which will turn out to have provided really important **new tools.** It is conspicuous that the revival of categorial grammar continues, that unification is a predominant procedure that government-binding designs are in vogue - or let me more respectfully say have proven fruitful - and that two-level models are used to such an extent and to a number of purposes which far exceeds mere politeness to our hosts. Tree-adjoining grammars seem to make some problems much more easy to treat.

Experiments with connectionist models are also promising; it would be unfair to require that they should be more than that.

4.2 The advances **beyond full stop** continue, though perhaps more laboriously than one might have hoped. Most attempts concern cohesion between neighbouring sentences - such as anaphoric phenomena - but interesting results have also been gained in generative form for larger structures (text planning). - Stylistic studies are rare as are other studies linking to literary theory. An attempt to identify text topics and one on concept analysis and terminology address issues directly relevant for document retrieval and topic 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, convenient user-interfaces at prices afordable even to front-line research institutes - have made experiments much more life-like. Performance and elegance as reported in the Project Notes are often impressive.

The ability to handle realistic vocabularies has led to a revived interest in lexicology. In particular, the knowledge accumulated in dictionaries for human use is being recycled by using these dictionaries as raw data supplies, which with some ingenuity are automatically exploited, or, to a smaller extent, as sources of know-how and linguistic insight.

While the translation tools are coming much closer to real life the pretensions have become more modest and more specific: systems are no longer designed to be omnivorous, when completed, but dedicated to particular kinds of texts and situations. Ironically, the success of such intelligent translation systems is almost cannibalistic: the most renowned successful computerized translation system, of weather reports, is reported to develop into a text generator. Is it so that the very mapping of the languages and the subject-matter which are required to make translation programmable also threatens to make it unnecessary?

4.5 Linguists have lifted their gaze to see **beyond the written text.** We catch a glimpse of the multimedia society in, say, a report on automatic generation of an animated presentation from a written instruction manual. On the whole, however, image processing is not yet part of computational linguistics.

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 reasonably with phenomena which are in some sense unexpected, except that the system was designed to expect 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, but 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 is the analysis of faulty input. Or let us say non-canonical input since it is nobody's fault 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 immaculate precision in human behaviour has also theoretical implications. Very serious implications, possibly: if there is more to it than can be handled by a little normalizing 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 remark: as one who spent a good deal of energy in earlier years to promote statistical methods in linguistics (and publishing a scientific journal, SMIL, for the purpose) I cannot but feel pleased at this development, 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 the quantitative studies in linguistics today is no higher than in statistical linguistics 30 years ago. Little efforts are spent on the non-trivial task of creating appropriate stochastic models for linguistics. And, quite unnecessarily, the accumulated knowhow of the statistical profession is disregarded: 'probability' and '(relative) frequency' are often used indiscriminately as synonyms, a discussion of estimates and sampling is mostly conspicuously absent and the reader is supplied with uninterpretable percentages.

This methodological weakness is probably due not only to an understandable unacquaintance 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, very many top-level mathematicians and logicians remain uninterested in stochastical models, unfamiliar with statistical practice and at bottom hostile towards the underlying attitude to look for a well-motivated approximation - to live with uncertainty instead of waiting for an ideal solution. We need more interdisciplinarity!

5. To characterize computational linguistics of 1990 we must also make a few remarks, necessarily 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 as a dozen highly idealized particles. The realistic experiments are more fully-fledged 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 in complexity and craving for fuller knowledge - and lack of the mathematician's delight in fruitful simplification. We have more barriers to break down: in this case the moral barrier 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 translation, 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 investigating 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. I could think of no other reason than the inherited sociological 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 better tools, has attracted no computational linguist this time, nor has authourship attribution. We have already remarked on the emptyness 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 more remarkable as phonetic 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 humanities! Which research could be more humanistic than examining the frontier between what in us is man and what in us is machine when we perform our most human activities?

Many papers contain a trace of bad conscience: we know we should deliver applicable results which justify what we cost now and if society only allows us time for basic preliminaries 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 manhours 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 greater than three, 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 manmonths, unpaid but highly rewarding, is a small trifle 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 whatever currency you choose to count in, and a great responsibility. These resources - which if spent for other causes in 1990 represent a power which could readily have rescued villagefuls of starving people from a painful death, bought thousands of minors free from being sent to sexual slavery, saved large areas of virgin forests from being forever devastated or bought some underprivileged country an extra submarine to fight its enemies - were supplied to us for a cause. They were not intended as subsidies to the computer 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.