# UNFINISHED LANGUAGE

## In the invitation to the electronic colloquium the topic was described in the following way:

No matter how comprehensively we linguists document language as is, language processing machines must sooner or later cope with unknown words, shades of meanings and concepts: human language is a language in the making, changing by small shifts during every text or dialogue.

Modeling language users' learning and adaptation 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.

Recent encouraging progress in handling very large data bases might obscure this crucial issue and postpone its solution.

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. Are the basic mechanisms common to language change over longer periods, to language acquisition by an individual and to 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?

Machine learning is being studied today by many means. But it is not unreasonable to 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.

Send your messages on this topic to Mail Moderator Walther von Hahn, vhahn@rz.informatik.uni-hamburg.dbp.de.

## MAIL MODERATOR'S SUMMARY

Walther v Hahn

University of Hamburg
Computer Science Department
email: vhahn@rz.informatik.uni-hamburg.bdp.de

In over 80 mail activities and several telefax mails the statements of about 15 contributors have been delivered and commented. The following text summarizes some of the contributions and adds further issues which in my view should be relevant for a panel discussion at Helsinki. Due to space limitations we cannot present all arguments with the same degree of elaboration. I apologize in advance for misinterpretations or too compressed statements.

### I. Learning

* The topic of 'Unfinished Language' has an **ontogenetic** and a **phylogenetic** side: The latter concerns the behaviour of a single system toward changes of the linguistic environment, the first is connected with historic aspects of linguistic material (and linguistic methods resp.) as a basis for system design and computatinal processes.

* Machine **learning** concerning **lexical entries** is an issue which came up in most statements and which arises since several years at Coling, because obviously the notion of a complete lexicon is rather absurd. On the other hand, research must always proceed as if the lexicon was complete. No practical progress is possible without this assumption at least for realistic tests. Every practical system has a lexical acquisition component; the maintenance of a system consists mostly in lexical updating (esp in documentation tools and machine translation) (Calzolari/Bindi).

Any updating of the lexicon or individual lexicon entries obviously requires to distinguish

(i) between missing entries (a lack of the present description) and updating (due to linguistic changes). In the first case a new entry adds something without deleting anything whereas in the second case an existing entry is affected and existing interrelations must be changed.

(ii) between input error, misspellings, misconceptions etc and new potential entries.

An appropriate treatment could be done interactively by the help of an expert system (Reimann).

* Machine **learning** concerning **syntax** is another topic (Hirschman), the discussion of which also touched on robust parsing and identification of syntactic islands in unparsible sentences (Reeker).

* Machine **learning** in general was addressed, being a field which gives no quick evidence for clear learning strategies, for sharp divisions of syntax, semantics, ontology etc.

* A superficial look at experiments with **artificial neural networks** might support the idea that a network will learn somehow 'automatically' without being guided by (psycho-)linguistic knowlege.Insiders are rather sceptical about mixing the lexical acquisition problem with neural network mechanisms (Schnelle). Hybrid systems, moreover, raise the problem of integrating inhomogenous symbolic representations (Wermter).

* Learning always raises the problem of **forgetting**: Assume a linguistic change has been detected. Which (parts of an) entry must be removed, which inferences must be withdrawn etc to maintain consistency?

### II. Language

* **Text analysis** can be regarded as a microcosmos of linguistic change (Haenelt/Koenyvas-Toth) because a text as a whole does not refer to one static concept or referent right from the beginning, but the text develops an incremental constitution of a meaning, starting with something definitely contained in the knowledge base and changing the semantic environment of the entry.

* Every linguistic system must allow meta-statements concerning **new semantic definitions** of existing words among the partners. From this moment onwards the new definition is valid and changes the language. This process is implicitly given in metaphorical use of language (Bateman).

* The lexical material even in a 5-year-project is linguistically **inhomogenous** because the difference between co-workers and because of the changing view of domain description. Especially, the words attached to a concept hierarchy will change with every enhancement of the domain or the linguistic coverage.

### III. Speaker/Hearer and World

* The most prominent area in which changes in time are relevant for everyday application is **speaker adaptation** in speech recognition. Every speaker must undergo periodical post-adaptation to the same speaker.

* The **Individual linguistic history** of the one who writes the linguistic model is a permanent reason of change. His/her intuition and focus of attention in principle changes permanently.

* It is one of the basic requirements of **user models** that the entries change with the ongoing dialogue. As it was shown in several publications, user models are tightly connected to linguistic dialogue history.

* Changes in the **subject domain** causes changes in the (reference) semantic description (Huang Jianshuo). More philosophically: The reason for Man-Machine Interaction is an unfinished action or plan, otherwise no communication would be necessary.

## IV. Method and Application

* Unfinished language from the point of view of lingistics: Language **description can never be complete**. The history of linguistics is evidence enough for this issue. Research in linguistic change requires a more psycholinguistic approach (derived from what we know of language acquisition) than grammarians will normally follow (Powers).

* Processing techniques in computational linguistics must be open enough. Ununderstood and unexpected phenomena can be modelled (not only captured in a preliminary procedure) by probabilistic methods (Schubert).

* Linguists will admit that all **models** of the language are **temporary**. We know that there is much discussion among the engineering wing of computational linguistics whether they should follow every second year another XYZ-grammar.

* Even new and more elaborated **processing strategies** (in being sensitive to linguistic theory) make natural langauge systems unfinished.

* Speaking about unfinished language among (computational) linguists as a principle will evoke approval and seems to be nearly trivial. On the other hand in real computation the changes of the data are more or less **neglectable** compared to the huge amount of open questions with existing systems (Melby).

* In descriptions of existing linguistic corpora the most prominent data attached to them seem to be size, text type and coverage but not the **'age'** of the material and the relative time span in which it was built up.

Addresses:

John A. Bateman (Marina del Rey, CA, USA)
e-mail: bateman@isl.edu

Nicoletta Calzolari/Remo Bindi (Pisa, Italy)
e-mail: glottolo@icnucevm.bitnet

Huang Jianshuo (Guangzhou, 510641 PR China)
South China University of Technology

Karin Hänelt/Michael Könyves-Toth (Darmstadt, Germany)
e-mail: haenelt@ipsi.darmstadt.gmd.dbp.de /
koenyves@ipsi.darmstadt.gmd.dbp.de

Lynette Hirschman (Paoli, PA, USA)
e-mail: hirsch@prc.unisys.com

Alan Melby (Provo, Utah, USA)
e-mail: melby@byuvm

David Powers (Kaiserslautern, Germany)
e-mail: powers@uklirb.uucp

Larry Reeker, Alexandria, VA, USA)
e-mail: reeker@ida.org

Dorothea Reimann (Berlin, DDR)
Prenzlauer Promenade 149-152

Klaus Schubert (Utrecht, Netherlands)
e-mail: schubert@dlt1.uucp

Stefan Wermter (Dortmund, Germany)
e-mail: wermter@unido.uucp