Mid term evaluation Spoken Dutch Corpus project

Conclusions and recommendations Evaluation Committee

Prof. dr. Reinier Salverda
University College London/Faculty of Arts & Humanities

Prof. dr. Steven Bird
University of Pennsylvania/Linguistic Data Consortium

Dr. Jan Hajiç
Charles University, Prague

Prof. dr. Harald Höge
Siemens AG, München

1. Overall conclusion of the scientific evaluation: “On the whole we feel that this is a truly excellent project, with a high standard of scholarship that will bring great benefits to the study of Dutch for many years to come.”
2. Overall conclusion of the technical evaluation: “The CGN project is an ambitious task which when finished will definitely boost the development of speech and language technology and encourage basic and applied research for Dutch and Flemish. Compared to validation results in SpeechDat projects and other comparable speech data collection efforts, the third release of the CGN corpus shows in both Dutch and Flemish components good to very good results.”
3. Compared to the former British National Corpus (BNC) project, which was a first attempt to build a corpus describing the national language, the CGN project is a big step forward in improving the usability for language research. In particular the CGN project makes available the speech signal itself allowing combined research on the acoustic and linguistic level. As such the CGN project is an excellent example for other national projects which will hopefully soon be undertaken, especially in Europe to preserve the cultural heritage of the European languages.
4. The CGN project is an example of successful multidisciplinary international cooperation by scholars from Flanders and the Netherlands. The funding of □ 4.7 million is sufficient to cover the basic needs of research by delivering a 10 million-word corpus. According to the progress so far reached it seems very probable that the ‘10 million goal’ - although very ambitious - will be reached. The groups involved in the design of the corpus i.e. speech technologists, linguists, phoneticians and sociolinguists from the Netherlands and Flanders, found a good compromise to cover their respective research needs given the limited funding. The project is within budget, and offers excellent value for money.
5. The CGN is setting standards internationally, and for this reason it should make all documentation and frozen protocols available in English on the web.
6. The CGN project has reached a status, where the cornerstones – the design of the corpus, the protocols and the basic recordings – have been set properly. Nevertheless a number of recommendations can be given for further improvement of the CGN project. Most of these concern follow up actions of the project. For the remaining project period, the CGN should consider in particular the more detailed technical and scientific recommendations listed below under 4 and 5. We also recommend a continuation of user evaluation and an additional validation for the syntactic annotations of the corpus as soon as they become available.
7. The CGN Board should start preparing for the future by developing ideas and plans for research projects in innovative language research that will use the CGN as a key resource.
8. For the maintenance and distribution of the CGN after October 2003, institutional arrangements have been made with the Dutch Language Union. Proper budgetary arrangements still need to be made.

General introduction and evaluation protocol

This report contains the results of the mid-term evaluation of the Spoken Dutch Corpus (CGN-project), which was held on 24-25 October 2001. The aim of this evaluation was to assess the scientific and technical quality of the work done during the first half (1998-2001) of the NWO-project Corpus Spoken Dutch (Corpus Gesproken Nederlands, CGN).
The evaluation itself consisted of two parts: a technical evaluation by the Bavarian Archive for Speech Signals (BAS, München), and a scientific evaluation by an international evaluation committee.

The international evaluation committee was composed of the following experts:

- Prof. dr. Reinier Salverda
  University College London/Faculty of Arts & Humanities, London, United Kingdom
- Prof. dr. Steven Bird
  University of Pennsylvania/Linguistic Data Consortium, Philadelphia, United States
- Dr. Jan Hajiç
  Charles University/Institute of Formal and Applied Linguistics, Prague, Czech Republic
- Prof-dr. Harald Höge
  Siemens AG, München, Germany

BAS performed the technical evaluation on the 3rd release of CGN during the summer of 2001. The final version of the BAS-report is included in the appendix, as well as some additional information about how the technical evaluation was performed. On behalf of BAS, dr. Ch. Draxler, responsible for the evaluation, was present during the evaluation.

The scientific evaluation took place on 24 October 2001 at the Max-Planck-Institute for Psycholinguistics in Nijmegen, the Netherlands. The programme of the day is included. Input to this part of the evaluation were the following documents which were made available to the members of the committee a few weeks beforehand:

1. General information about the CGN-project as presented on the CGN-website (http://lands.let.kun.nl/cgn/ehome.htm), in Dutch;
2. The technical evaluation report, i.e. the BAS-report;
3. Eight CGN documents:
   a) description of the meta-data
   b) description of the orthographic transcription
   c) description of the Part-of-Speech tagging and lemmatization
   d) description of the sound files
   e) description of the XML-format
   f) motivation of the protocol for the orthographic transcription
   g) procedure for the orthographic transcription
   h) procedure for the Part-of-Speech tagging and lemmatisation;
4. A series of conference papers;
5. The annotation protocols (in Dutch, since they are to be used by annotators) were also available for inspection on the website;
6. A set of evaluation criteria.

The members of the evaluation committee were asked to formulate their comments on the basis of the documentation, and to present them following the criteria (document 6). They responded to those questions they felt lying in their specific field of expertise. These comments were gathered and distributed prior to the actual evaluation, and served as a guideline for the discussions on the 24th. These comments, completed together with the outcome of the discussions of the 24th, form the basis of the present report.

On October 24th, the committee first met in a closed session with a delegation of the board of CGN. On behalf of the board were present: prof. dr. W.J.M. Levelt (chairman), prof. dr. G. Booij, prof. dr. S.G. Nooteboom, prof. dr. D. Van Compernolle and dr. C. Cucchiarini (on behalf of the Dutch Language Union). This session was followed by a plenary discussion with a delegation of the steering committee. On behalf of the steering committee were present: prof. dr. J.-P. Martens (chairman, Ghent), dr. H. Baayen (Nijmegen), prof. dr. S. Gillis (Antwerpen), dr. H. van den Heuvel (Nijmegen), prof. dr. R. van Hout (Nijmegen/Tilburg), dr. J. Odijk (Lernout & Hauspie), dr. N. Oostdijk (Nijmegen), prof. dr. ir. L.C.W. Pols (Amsterdam), dr. I. Schuurman (Leuven) and dr. A. van der Wouden (Utrecht). The first half of the discussion concentrated on the scientific part of the evaluation, the second half on the technical part, i.e. the validation by BAS.
At the end of the afternoon, two members of the user group were invited for discussion. Present were dr. L. ten Bosch (Lernout & Hauspie) and dr. T. Kruyt (Institute for Dutch Lexicology).
The evaluation committee was supported by dr. M. van Donzel and drs. A. Dijkstra of NWO.

On the 25th, an open symposium was organized, to inform all interested in the project about the outcome of the evaluation. The morning session was dedicated to this topic. In the afternoon, presentations were given by Harald Höge on current research activities at Siemens, by Martin Wynne of the British National Corpus, and Christoph Draxler on how the technical evaluation of CGN was performed. This last presentation was the link back to the final plenary discussion on the recommendations and conclusions of the evaluation.

Assessment criteria scientific evaluation[1]

General

1. The timeframe of the CGN-project including the design phase is 5 years. Is this realistic? Considering that the project started in the summer of 1998 and will finish in the summer of 2003, do you think the project, at the moment of this mid-term evaluation, is indeed halfway?
2. The total budget for the CGN-project is □ 4.7 million (approximately 4 million US$). With the goal of the project in mind, would you consider this expensive/low budget/appropriate?

Design and Annotation choices

3. Is the overall design of the corpus appropriate for its goal? Is the distribution of speech material (dialogue vs. monologue, scripted vs. non-scripted speech etc.) logically chosen? Is the selection of data for which more advanced transcriptions and annotations are envisaged appropriate? Do you agree with the way the data is stored (uncompressed 16 bit, 16 kHz wav format)?
4. Do you think the choices made in the protocols are acceptable? Are the choices made properly motivated?
    * *Orthographic transcription* yes/no Comments?
    * *Part-of-Speech tagging* yes/no Comments?
    * *Syntactic transcription* yes/no Comments?

Conclusions of the technical evaluation performed by BAS

5. Which BAS' conclusions should lead to which CGN actions?
6. Which conclusions are important, but cannot be remedied anymore at this stage of the project? Does it have any critical consequences for the final results?

Internationalisation

7. Are the institutions that are responsible for designing and building the CGN aware of the international state of the art? Can the project be considered state-of-the-art? What do you consider as shortcomings in this respect?
8. Is the international community aware of the CGN-project?

Overall

9. What is your general opinion of the CGN-project in terms of plans, motivation, technical aspects, execution, etc.? In which ways can it be considered exemplary, in which ways not? CGN learnt a lot from the British National Corpus project (BNC); what can the international community learn from the CGN-project?
10. Are there any critical gaps in the project? Do you have any suggestions for improvement?

---

[1] The development of exploration software (COREX) is part of the CGN-project. The first version of the COREX-software has just been delivered and is now being tested by the user group. Assessing the COREX-software is not part of the midterm evaluation.

Specific comments to the assessment criteria

Introduction

- The evaluation committee has a very positive view of the quality and the progress of the CGN-project: "On the whole [the committee] feel[s] that this is a truly excellent project, with a high standard of scholarship that will bring great benefits to the study of Dutch for many years to come."
  The significance of the CGN-project is that it can be compared to the building of the high velocity particle accelerator at CERN in Geneva, but a lot cheaper. That is, CGN aims to construct a major tool for innovative research in language. In this respect, the two key aims of the project are (i) to contribute to the advancement of Dutch and Flemish speech and language technology, and (ii) to address the needs of lexicographers, linguists, psycholinguists, conversational analysts and curricula developers in the Dutch speaking Low Countries. While the CGN-project has a number of wider aims in the field of education, business communication and the politics of language, these are less central to the construction stage of the project. They were therefore not the first concern of the evaluation committee.

  The corpus that is now under construction is a vast thing to produce, whereas scientific progress often requires specialization. The key virtue of the CGN-project is its successful multi-disciplinary cooperation, in a broad framework that brings together speech technologists, linguists, phoneticians and sociolinguists from the Netherlands and Flanders, all working together effectively towards the two central aims of the project.

General

1. *The timeframe of the CGN-project, including the design phase, is 5 years. Is this realistic? Considering that the project started in the summer of 1998 and will finish in the summer of 2003, do you think the project, at the moment of this mid-term evaluation, is indeed halfway?*

- Considering the size of the corpus (10 million words), the timeframe was realistically chosen. The evaluation committee noted that delays have occurred, especially in the initial phase in 1998. More time than initially anticipated was spent on the pilot project for the development of protocols, and on the effective organization, negotiations and compromises within the multidisciplinary CGN-team. Also, the development of the protocols proved a complex and challenging task, which has taken more time than originally expected. The evaluation committee considers that the time taken up must be regarded a sensible choice and a valuable investment.

  So far, four releases have been published [on 110 CDs], and another three releases are planned for the second phase of the project. In this respect, the CGN-project is now at the half-time mark, but a lot of data still has to be collected, transcribed and annotated. The committee urges the steering committee to continue their efforts to maintain as high a quality standard as possible for the corpus. The committee stresses that if choices must be made, a lesser number of words is preferred – while keeping the design valid - to lowering the excellent quality of the annotations, the protocols and tools.

  The committee would like to stress that a corpus is never finished, and for this reason the steering committee and board must start planning ahead, to prepare for possible extensions to the corpus and for follow-up research projects. The following kinds of extensions have been identified as desirable: children's speech, non-native Dutch, recordings in noisy and mobile environments, and the grammar of conversational Dutch.

2. *The total budget for the CGN-project is ☐ 4.7 million (approximately 4 million US$). With the goal of the project in mind, would you consider this expensive/low budget/appropriate?*

   - The budget seems adequate and appropriate for the choices made (c.f. design choices under 3 below). The committee has not found any indication of misspending of money within the CGN-project. It is a big project that offers value for money.
     The size of the corpus – 10 million words - was dictated by the size of the budget. Budgetary considerations have played a role in the corpus design, especially in the selection of 1 million words to be annotated, i.e. the so-called nuclear corpus. The committee stresses that future extensions with a) more annotations and b) other specific types of data, e.g. children's speech data, non native speech data, data for speech synthesis purposes and mobile telephone data is still required to satisfy both scientific and industrial users.
     The committee understands that the CGN steering committee after careful consideration and extensive discussions has chosen to invest especially in syntactic annotation and word alignment. Syntactic annotation is an expensive effort – mainly because even though automatic help can be provided, mostly it requires manual labour. Having a syntactically annotated corpus will allow for fascinating new research opportunities. As for word alignment, the steering committee has opted for the development of automatic alignment software which will allow CGN not only to provide a verified alignment for the nuclear corpus, but moreover an automatically produced alignment of the complete corpus. This word alignment, though not verified, is expected to have a more than reasonable quality, allowing better access to the complete corpus.

Design and Annotation choices

3. *Is the overall design of the corpus appropriate for its goal? Is the distribution of speech material (dialogue vs. monologue, scripted vs. non-scripted speech etc.) logically chosen? Is the selection of data for which more advanced transcriptions and annotations are envisaged appropriate? Do you agree with the way the data is stored (uncompressed 16 bit, 16 kHz wav format)?*

   - Preliminary: goals of the project
     The CGN project aims at collecting, transcribing and (partly) annotating 10 million words of spoken Dutch. Such a corpus is essential for two types of users:
     * as a basis for the development of Dutch automatic speech recognition (ASR) software;
     * as a basis for research of lexicographers, linguists, psycholinguists, conversational analysts and curricula developers.

     The choices made within the CGN project with respect to the design – both as to its components,its contents and its various levels of annotation - reflect the multiple goals of the project. Considering the two different types of users, sensible compromises, especially with respect to the choices made in defining the design of the corpus were required. Experts from both types are part of the steering committee and represented in the working group responsible for the final design.
     The evaluation committee understands that the CGN design is fitted both to the technological and the scientific goal. It is a huge step forward in setting criteria, guidelines, protocols and tools for creating these kinds of linguistic resources. Ample opportunities for corpus extensions for specific needs of both types of users are thus provided.
     The evaluation committee would appreciate if the motivation written for the corpus design were to more explicitly explain the compromises made. The CGN claims it is a 'plausible sample', and the evaluation committee does not contest this, but should like to see further clarification on this point - what is the CGN a sample of, and why is it considered plausible? The committee furthermore suggests adding an overview of the semantic topics so far covered by the recordings. New recordings should be selected towards enrichment of semantic coverage.

- Format

  As to the format of the data the committee urges the CGN to continue opting for the highest possible level of accuracy. Furthermore, to allow for ease of use and future extensions all annotation levels should be properly linked. The documentation presented to the committee and the discussions with the CGN steering committee convinced the evaluation committee that this is also the goal of CGN. A few low-level but important issues remain:
  - ✓ The format of sound files is standard, however format conversion tools should be supplied.
  - ✓ The format of the metadata was not reported. It is important that it be provided in a machine-readable, implementation-neutral format (not just a private data format used by the "COREX" software).
  - ✓ The XML format for primary data, signal coupling data and POS-tags is critical for the reusability of the data. The design is good, with several strong points such as the use of XML, the ubiquitous identifiers, and the detailed information about the nature and reliability of the temporal anchors. Formal specifications of the DTDs and the internal structure and management of the identifiers still need some more attention in order to make the corpus easy to use and manage in the future.

4. *Do you think the choices made in the protocols are acceptable? Are the choices made properly motivated?*

- The general strategy is sensible: to follow and use, wherever available and adequate, the international encoding standards; to cooperate wherever possible with other similar projects; and to develop new protocols where necessary.

  Orthographic transcription

  The protocol for orthographic transcription looks very solid. The main concern is with the embedded annotations, like *d. CHILDES is cited as a source for this practice, yet CHILDES is developing a new format precisely because of the myriad problems caused by embedding annotations in the orthographic string. A better approach, the committee believes, is to represent such annotations in an analogous way to the .tag data, using standoff mark-up represented as XML attributes.

  The BAS evaluation shows that the overall quality of the orthographic transcription made on the basis of this protocol is very good.

  Part-of-Speech tagging

  The POS tagset is being developed in line with a long tradition of descriptive standard grammars in the Netherlands. Moreover, the authors have compared their work to the EAGLES standard, and adequately explained the differences.

  The TIMBL Combiner's accuracy (94.2 > 95.6%) – the automatic POS-tagger - is exceptionally good. As a result manual corrections are hardly necessary, reducing the cost of this task (☐ 0.03/word) immensely. The CGN-project is able to do so because there is much expertise available in this area. The BAS evaluation furthermore shows that only a – extremely low - 0.03 error rate was found.

  Syntactic Annotation

  The annotation guidelines are reasonably detailed, with many examples provided. The committee supports the decision to use the dependency syntax as the basis for the annotation, on top of the phrasal structure.

  The actual annotation process has just entered the production phase. An adapted version of Annotate software is used to produce a first automatic annotation which subsequently needs manual verification. The committee recommends that as soon as a sufficient amount of data is syntactically annotated, these too are technically evaluated by an independent institute such as BAS. (The same holds for the phonetic annotations the CGN project has just started to make.)

Conclusions of the technical evaluation performed by BAS

5. *Which BAS' conclusions should lead to which CGN actions?*

- BAS has performed a detailed low-level technical evaluation. On the basis of the validation report of BAS, the evaluation committee concludes that the Dutch/Flemish team creating the language resource did an excellent job! The evaluation committee subscribes to the general conclusion of the BAS report, viz:
  "The CGN project is an ambitious task which when finished will definitely boost the development of speech and language technology and encourage basic and applied research for Dutch and Flemish. Compared to validation results in SpeechDat projects and other comparable speech data collection efforts, the third release of the CGN corpus shows in both Dutch and Flemish components good to very good results. Most of the errors found in the formal validation can be easily corrected; no errors were found that exclude the usage of certain recordings."

  The evaluation committee urges CGN to consider especially the following BAS recommendations:
  - ✓ *Publish all specifications, including a priori corpus specifications, recording protocols and a posteriori log files describing deviations from the plan.*
  - ✓ *Consider future extensions to the corpus. The experience from Verbmobil and SmartKom shows that the value of a corpus multiplies with the number of additional representation or annotation levels. Therefore, provide a flexible multi-tier annotation interchange format that also allows non time-aligned representations that can be linked to time-aligned representations. Establish guidelines for making corrections and re-annotations.*
  - ✓ *Perform automatic procedures and checks wherever possible, e.g. format checkers for signal files, site analysis tools for HTML structures, a validating parser for XML files, perl scripts for the generation of frequency lists, etc. These tools and scripts ideally should be platform independent and be provided together with the corpus.*
  - ✓ *The consistency issue is of particular importance for files that are used in further processing steps, e.g. the orthographic annotation in Praat files which are the basis of all frequency counts, phonetic segmentations, and POS-tagging.*
  - ✓ *Make releases available to partners really working with the data immediately to find errors that cannot be detected by automatic checkers. Install a bug reporting mechanism and encourage reporting bugs, e.g. with a reward for every bug found.*
  - ✓ *Annotations change, signals don't. Hence it makes sense to provide the annotations on-line in the most recent version. It is not clear whether a version control system is implementable for corpus data, but every result obtained must be accompanied by the version (version number, date released, source, etc.) of the data used.*
  - ✓ *Add a detailed documentation about the technical set-up of the recordings. For instance in recordings of radio shows it would be interesting to know about the exact procedure of the recording: master signal of the broadcasting station, recording via digital broadcasting (which type of receiver, converter, etc.), recording via analog receivers etc.*

6. *Which conclusions are important, but cannot be remedied anymore at this stage of the project? Does it have any critical consequences for the final results?*

- The evaluation committee is convinced that all of the recommendations made by BAS are feasible; there are no critical consequences for the final results.

Internationalisation

7. *Are the institutions that are responsible for designing and building the CGN aware of the international state of the art? Can the project be considered state-of-the-art? What do you consider as shortcomings in this respect?*

- The CGN-partners are certainly well aware of the international state of the art in the field and follow the latest developments in linguistically annotated corpora around the world (Penn Treebank version 3, Italian ILR corpus, Czech PDT 1.0), at least for reference purposes. In fact the evaluation committee believes that CGN has already moved beyond that and is setting new standards itself.
  Compared to the former British National Corpus (BNC) project, which was a first attempt to build a corpus describing the national language, the CGN project is a big step forward in improving the usability for language research. In particular the CGN project makes available the speech signal itself allowing combined research on the acoustic and linguistic level. The BNC- Corpus is however much larger, i.e. a 100 million word collection of samples of both spoken and written current British English but lacks on the availability of the speech signal of the spoken material. As such the CGN project is an excellent example for other national projects which will hopefully soon be undertaken, especially in Europe to preserve the cultural heritage of the European languages. Considering the quality of the standards the CGN project is defining, the evaluation committee would like to encourage CGN to make all documentation and frozen protocols available in English on the web as this will be instructive for those who would like to develop electronic speech and language corpora in other countries and for other languages.

8. *Is the international community aware of the CGN-project?*

- The evaluation committee truly appreciates the effort that has been made to further the knowledge transfer from the CGN project to future users of the corpus and to the international scientific community at large.
  Two user workshops have been organised and one international workshop on the design, compilation and annotation of the CGN. Furthermore, project members have produced an impressive list of national (14) and international (8) publications, and national (17) and international (10) presentations. The international publications generated by the CGN-project and the international distribution of the corpus via ELRA [the European Language Resources Agency] will do much to make the CGN widely known internationally. The international profile of CGN could be enhanced if free samples of data were published, and if all format definitions were made available so that software developers could support CGN formats.

Overall

9. *What is your general opinion of the CGN-project in terms of plans, motivation, technical aspects, execution, etc.? In which ways can it be considered exemplary, in which ways not? CGN learnt a lot from the British National Corpus project (BNC); what can the international community learn form the CGN-project?*

- The evaluation committee has a very high opinion of the quality and progress of the CGN-project. The protocols and software produced by the CGN-project constitute important and innovative tools which greatly enhance the usability of the corpus. No other such high quality corpus of spoken language exists anywhere. In these respects, the CGN is setting the standard internationally.
  The CGN project is a truly multidisciplinary project where a Dutch/Flemish team of speech technologists, linguists and phoneticians are successfully co-operating to create a valuable language resource that will bring great benefit to the study of Dutch for many years to come and provide an invaluable basis for the development of language engineering technologies. The evaluation committee sees the CGN-project as an exemplary model for Dutch-Flemish scientific cooperation in the field of basic language research.

10. Are there any critical gaps in the project? Do you have any suggestions for improvement?

- The committee has not identified any critical gaps in the CGN-project. It has identified a number of weaknesses that can be remedied over the remaining project period:
  - ✓ continue the technical evaluation and validation effort; step up user evaluation now that sufficient material is available; include both scientific and industrial users;
  - ✓ provide complete and detailed documentation and publish all specifications and protocols in English so that others can learn from CGN;
  - ✓ continue to strive for consistency; use one common distribution (data interchange) format; XML, with possible incorporation of TEI (P4X) whenever possible. The automatic tools performing the consistency checks should be platform independent and be made available together with the corpus;
  - ✓ consider future extensions and corrections to the corpus by providing a flexible multi-tier annotation interchange format, (cross) link all levels of annotations and provide guidelines for making corrections and re-annotations.

Future arrangements

The evaluation committee is satisfied that the CGN board has made institutional arrangements with the Dutch Language Union in The Hague for the maintenance and distribution of the CGN-corpus after the end of this project in October 2003. However, the evaluation committee notes with concern that there is no certainty about the financial aspects of these arrangements. Considering that for many years to come a sum of ƒ 1 million per year may be necessary for the proper management of the CGN, the evaluation team recommends that proper budgetary preparations are made.

Apart from this, there is the matter of further development of the corpus, and - most importantly - its use as a key resource for innovative research in language. The evaluation team recommends that the CGN board and steering committee start preparing for the future by developing ideas and plans for such future research as well as applications for funding of specific research projects.

Short curriculum vitae of members evaluation committee

Reinier Salverda (chairman)
University College London, London
Studied Dutch language and literature at the Free University of Amsterdam, where he obtained his PhD in Linguistics in 1985. From 1981 to 1989 he taught Dutch and Linguistics at the Universitas Indonesia in Jakarta. Since 1989 he is Professor of Dutch Language and Literature at University College London. His linguistic publications are focussed on the grammar of modern standard Dutch and the history of linguistics. He is a member of the Advisory Board of the Dutch Standard Grammar (ANS, Algemene Nederlandse Spraakkunst), and is currently preparing an Internet Grammar of Dutch.

Steven Bird
University of Pennsylvania, Philadelphia
Did his doctoral and post-doctoral research in Scotland at the University of Edinburgh, Centre for Cognitive Science (1987-94). From 1995-1997 he conducted linguistic fieldwork on the tone languages of western Cameroon, and helped develop several new writing systems. His undergraduate and postgraduate training were at the University of Melbourne, Department of Computer Science (1982-1987).
As of 1998 he is research scientist and associate director at the Linguistic Data Consortium, University of Pennsylvania, USA, as well as adjunct associate professor at the Department of Computer and Information Science, Department of Linguistics, University of Pennsylvania, USA.

Jan Hajiç
Charles University, Prague
Obtained his PhD in „Mathematical Linguistics“ from Charles University, Prague, in 1994, in Czech Morphology. Until 1991 he worked at the Institute of Mathematical Machines in Prague on a machine translation project from Czech to Russian, then at IBM T.J.Watson Research Center in Yorktown Heights, NY, USA, on a statistically-based machine translation project (French to English). Since 1993 he is an Assistant Professor at Charles University, Prague, working on lexicons, POS and morphological tagging of inflective languages, and he supervises the creation of a (deeply) syntactically annotated treebank of Czech. On his sabbatical in 1999-2000 he taught Natural Language Processing and worked in the Language and Speech area at the Computer Science Department and the Center of Language and Speech Procesing at the Johns Hopkins University, Baltimore, MD, USA.

Harald Höge
Siemens, München
Harald Höge received his diploma on physics in 1970 and his Ph.D. in 1974 from the university of Frankfurt/Main Germany.
1970 he joint the Siemens AG working on echo compensation and speech coding. Since 1978 he leads a research group focusing on speech recognition and recently on speech synthesis. He initiated and was involved in many national, European and international project as SPICOS, C-STAR, Verbmobil, SpeechDat, SALA, SPEECON and LC-STAR. He gives lectures at the 'Universität der Bundeswehr München' on speech and image processing, where he was honoured in 2001 with the title 'Honorar Professor'. He helped substantially to set up a speech processing group at University of Maribor (Slovene), the German language resource centre BAS at university of Munic and the European Language Resource Association ELRA, where he is active in the Board. He is author and co-author of 55 papers and holds 20 patents.