

# Natural Language Processing at the School of Information Studies for Africa

**Björn Gambäck**

Userware Laboratory  
Swedish Institute of Computer Science  
Box 1263, SE-164 29 Kista, Sweden  
gamback@sics.se

**Gunnar Eriksson**

Department of Linguistics  
Stockholm University  
SE-106 91 Stockholm, Sweden  
gunnar@ling.su.se

**Athanassia Fourla**

Swedish Program for ICT in Developing Regions  
Royal Institute of Technology/KTH  
Forum 100, SE-164 40 Kista, Sweden  
afourla@dsv.su.se

## Abstract

The lack of persons trained in computational linguistic methods is a severe obstacle to making the Internet and computers accessible to people all over the world in their own languages. The paper discusses the experiences of designing and teaching an introductory course in Natural Language Processing to graduate computer science students at Addis Ababa University, Ethiopia, in order to initiate the education of computational linguists in the Horn of Africa region.

## 1 Introduction

The development of tools and methods for language processing has so far concentrated on a fairly small number of languages and mainly on the ones used in the industrial part of the world. However, there is a potentially even larger need for investigating the application of computational linguistic methods to the languages of the developing countries: the number of computer and Internet users of these countries is growing, while most people do not speak the European and East-Asian languages that the computational linguistic community has so far mainly concentrated on. Thus there is an obvious need to develop a wide range of applications in vernacular languages, such as translation systems, spelling and grammar checkers, speech synthesis and recognition, information retrieval and filtering, and so forth.

But who will develop those systems? A prerequisite to the creation of NLP applications is the education and training of computer professionals skilled in localisation and development of language processing resources. To this end, the authors were invited to conduct a course in Natural Language Processing at the School of Information Studies for Africa, Addis Ababa University, Ethiopia. As far as we know, this was the first computational linguistics course given in Ethiopia and in the entire Horn of Africa region.

There are several obstacles to progress in language processing for new languages. Firstly, the particulars of a language itself might force new strategies to be developed. Secondly, the lack of already available language processing resources and tools creates a vicious circle: having resources makes producing resources easier, but not having resources makes the creation and testing of new ones more difficult and time-consuming.

Thirdly, there is often a disturbing lack of interest (and understanding) of the needs of people to be able to use their own language in computer applications — a lack of interest in the surrounding world, but also sometimes even in the countries where a language is used (“Aren’t those languages going to be extinct in 50–100 years anyhow?” and “Our company language is English” are common comments).

And finally, we have the problem that the course described in this paper mainly tries to address, the lack of skilled professionals and researchers with knowledge both of language processing techniques and of the domestic language(s) in question.

The rest of the paper is laid out as follows: The next section discusses the language situation in Ethiopia and some of the challenges facing those trying to introduce NLP methods in the country. Section 3 gives the background of the students and the university, before Section 4 goes into the effects these factors had on the way the course was designed.

The sections thereafter describe the actual course content, with Section 5 being devoted to the lectures of the first half of the course, on general linguistics and word level processing; Section 6 is on the second set of lectures, on higher level processing and applications; while Section 7 is on the hands-on exercises we developed. The evaluation of the course and of the students' performance is the topic of Section 8, and Section 9 sums up the experiences and novelties of the course and the effects it has so far had on introducing NLP in Ethiopia.

## 2 Languages and NLP in Ethiopia

Ethiopia was the only African country that managed to avoid being colonised during the big European power struggles over the continent during the 19th century. While the languages of the former colonial powers dominate the higher educational system and government in many other countries, it would thus be reasonable to assume that Ethiopia would have been using a vernacular language for these purposes. However, this is not the case. After the removal of the Dergue junta, the Constitution of 1994 divided Ethiopia into nine fairly independent regions, each with its own "nationality language", but with Amharic being the language for countrywide communication. Until 1994, Amharic was also the principal language of literature and medium of instruction in primary and secondary schools, but higher education in Ethiopia has all the time been carried out in English (Bloor and Tamrat, 1996).

The reason for adopting English as the *Lingua Franca* of higher education is primarily the linguistic diversity of the country (and partially also an effect of the fact that British troops liberated Ethiopia from a brief Italian occupation during the Second World War). With some 70 million inhabitants, Ethiopia is the third most populous African country and harbours more than 80 different languages — exactly how many languages there are in a country

is as much a political as a linguistic issue; the count of languages of Ethiopia and Eritrea together thus differs from 70 up to 420, depending on the source; with, for example, the Ethnologue (Gordon, 2005) listing 89 different ones.

Half-a-dozen languages have more than 1 million speakers in Ethiopia; three of these are dominant: the language with most speakers today is probably Oromo, a Cushitic language spoken in the south and central parts of the country and written using the Latin alphabet. However, Oromo has not reached the same political status as the two large Semitic languages Tigrinya and Amharic. Tigrinya is spoken in Northern Ethiopia and is the official language of neighbouring Eritrea; Amharic is spoken in most parts of the country, but predominantly in the Eastern, Western, and Central regions. Oromo and Amharic are probably two of the five largest languages on the continent; however, with the dramatic population size changes in many African countries in recent years, this is difficult to determine: Amharic is estimated to be the mother tongue of more than 17 million people, with at least an additional 5 million second language speakers.

As Semitic languages, Amharic and Tigrinya are distantly related to Arabic and Hebrew; the two languages themselves are probably about as close as are Spanish and Portuguese (Bloor, 1995). Speakers of Amharic and Tigrinya are mainly Orthodox Christians and the languages draw common roots to the ecclesiastic Ge'ez still used by the Coptic Church. Both languages use the Ge'ez (Ethiopic) script, written horizontally and left-to-right (in contrast to many other Semitic languages). Written Ge'ez can be traced back to at least the 4th century A.D. The first versions of the script included consonants only, while the characters in later versions represent consonant-vowel pairs. Modern Amharic words have consonantal roots with vowel variation expressing difference in interpretation.

Several computer fonts have been developed for the Ethiopic script, but for many years the languages had no standardised computer representation. An international standard for the script was agreed on only in year 1998 and later incorporated into Unicode, but nationally there are still about 30 different "standards" for the script, making localisation of language processing systems and digital resources

difficult; and even though much digital information is now being produced in Ethiopia, no deep-rooted culture of information exchange and dissemination has been established. In addition to the digital divide, several other factors have contributed to this situation, including lack of library facilities and central resource sites, inadequate resources for digital production of journals and books, and poor documentation and archive collections. The difficulties of accessing information have led to low expectations and consequently under-utilisation of existing information resources (Furzey, 1996).

UNESCO (2001) classifies Ethiopia among the countries with “moribund or seriously endangered tongues”. However, the dominating languages of the country are not under immediate threat, and serious efforts have been made in the last years to build and maintain linguistic resources in Amharic: a lot of work has been carried out mainly by Ethiopian Telecom, Ethiopian Science and Technology Commission and Addis Ababa University, as well as by Ethiopian students abroad, in particular in Germany, Sweden and the United States. Except for some initial efforts for the related Tigrinya, work on other Ethiopian languages has so far been scarce or non-existent — see Alemu et al. (2003) or Eyassu and Gambäck (2005) for short overviews of the efforts that have been made to date to develop language processing tools for Amharic.

One of the reasons for fostering research in language processing in Ethiopia was that the expertise of a pool of researchers in the country would contribute to maintaining those Ethiopian languages that are in danger of extinction today. Starting with Amharic and developing a robust linguistic resource base in the country, together with including the Amharic language in modern language processing tools could create the critical mass of experience, which is necessary in order to expand to other vernacular languages, too.

Moreover, the development of those conditions that lay the foundations for language and speech processing research and development in the country would prevent potential brain drain from Ethiopia; instead of most language processing work being done by Ethiopian students abroad (at present), in the future it could be done by students, researchers and professionals inside the country itself.

### 3 Infrastructure and Student Body

Addis Ababa University (AAU) is Ethiopia’s oldest, largest and most prestigious university. The Department of Information Science (formerly School of Information Studies for Africa) at the Faculty of Informatics conducts a two-year Master’s Program. The students admitted to the program come from all over the country and have fairly diverse backgrounds. All have a four-year undergraduate degree, but not necessarily in any computer science-related subject. However, most of the students have been working with computers for some time after their under-graduate studies. Those admitted to the program are mostly among the top students of Ethiopia, but some places are reserved for public employees.

The initiative of organising a language processing course as part of the Master’s Program came from the students themselves: several students expressed interest in writing theses on speech and language subjects, but the faculty acknowledged that there was a severe lack of staff qualified to teach the course. In fact, all of the university is under-staffed, while admittance to the different graduate programs has been growing at an enormous speed; by 400% only in the last two years. There was already an ICT support program in effect between AAU and SAREC, the Department for Research Cooperation at the Swedish International Development Cooperation Agency. This cooperation was used to establish contacts with Stockholm University and the Swedish Institute of Computer Science, that both had experience in developing computational linguistic courses.

Information Science is a modern department with contemporary technology. It has two computer labs with PCs having Internet access and lecture rooms with all necessary aids. A library supports the teaching work and is accessible both to students and staff. The only technical problems encountered arose from the frequent power failures in the country that created difficulties in teaching and/or loss of data. Internet access in the region is also often slow and unreliable. However, as a result of the SAREC ICT support program, AAU is equipped with both an internal network and with broadband connection to the outside world. The central computer facilities are protected from power failures by generators, but the individual departments have no such back-up.

## 4 Course Design

The main aim of the course plan was to introduce the students successfully to the main subjects of language and speech processing and trigger their interest in further investigation. Several factors were important when choosing the course materials and deciding on the content and order of the lectures and exercises, in particular the fact that the students did not have a solid background in either Computer Science or Linguistics, and the time limitations as the course could only last for ten weeks. As a result, a curriculum with a holistic view of NLP was built in the form of a “crash course” (with many lectures and labs per week, often having to use Saturdays too) aiming at giving as much knowledge as possible in a very short time.

The course was designed before the team travelled to Ethiopia, but was fine-tuned in the field based on the day-by-day experience and interaction with the students: even though the lecturers had some knowledge of the background and competence of the students, they obviously would have to be flexible and able to adjust the course set-up, paying attention both to the specific background knowledge of the students and to the students’ particular interests and expectations on the course.

From the outset, it was clear that, for example, very high programming skills could not be taken for granted, as given that this is not in itself a requirement for being admitted to the Master’s Program. On the other hand, it was also clear that *some* such knowledge could be expected, this course would be the last of the program, just before the students were to start working on their theses; and several laboratory exercises were developed to give the students hands-on NLP experience.

Coming to a department as external lecturers is also in general tricky and makes it more difficult to know what actual student skill level to expect. The lecturer team had quite extensive previous experiences of giving external courses this way (in Sweden and Finland) and thus knew that “the home department” often tends to over-estimate the knowledge of their students; another good reason for trying to be as flexible as possible in the course design. and for listening carefully to the feedback from the students during the course.

The need for flexibility was, however, somewhat counter-acted by the long geographical distance and time constraints. It was necessary to give the course in about two months time only, and with one of the lecturers present during the first half of the course and the other two during the second half, with some overlap in the middle. Thus the course was split into two main parts, the first concentrating on general linguistic issues, morphology and lexicology, and the second on syntax, semantics and application areas.

The choice of reading was influenced by the need not to assume very elaborated student programming skills. This ruled out books based mainly on programming exercises, such as Pereira and Shieber (1987) and Gazdar and Mellish (1989), and it was decided to use Jurafsky and Martin (2000) as the main text of the course. The extensive web page provided by those authors was also a factor, since it could not be assumed that the students would have full-time access to the actual course book itself. The costs of buying a regular computer science book is normally too high for the average Ethiopian student.

To partially ease the financial burden on the students, we brought some copies of the book with us and made those available at the department library. We also tried to make sure that as much as possible of the course material was available on the web. In addition to the course book we used articles on specific lecture topics particularly material on Amharic, for which we also created a web page devoted to online Amharic resources and publications.

The following sections briefly describe the different parts of the course and the laboratory exercises. The course web page contains the complete course materials including the slides from the lectures and the resources and programs used for the exercises:

[www.sics.se/humle/ile/kurser/Addis](http://www.sics.se/humle/ile/kurser/Addis)

## 5 Linguistics and word level processing

The aim of the first part of the course was to give the students a brief introduction to Linguistics and human languages, and to introduce common methods to access, manipulate, and analyse language data at the word and phrase levels. In total, this part consisted of seven lectures that were accompanied by three hands-on exercises in the computer laboratory.

## 5.1 Languages: particularities and structure

The first two lectures presented the concept of a human language. The lectures focused around five questions: What is language? What is the ecological situation of the world's languages and of the main languages of Ethiopia? What differences are there between languages? What makes spoken and written modalities of language different? How are human languages built up?

The second lecture concluded with a discussion of what information you would need to build a certain NLP application for a language such as Amharic.

## 5.2 Phonology and writing systems

Phonology and writing systems were addressed in a lecture focusing on the differences between writing systems. The SERA standard for transliterating Ethiopic script into Latin characters was presented. These problems were also discussed in a lab class.

## 5.3 Morphology

After a presentation of general morphological concepts, the students were given an introduction to the morphology of Amharic. As a means of handling morphology, regular languages/expressions and finite-state methods were presented and their limitations when processing non-agglutinative morphology were discussed. The corresponding lab exercise aimed at describing Amharic noun morphology using regular expressions.

In all, the areas of phonology and morphology were allotted two lectures and about five lab classes.

## 5.4 Words, phrases and POS-tagging

Under this heading the students were acquainted with word level phenomena during two lectures. Tokenisation problems were discussed and the concept of dependency relations introduced. This led on to the introduction of the phrase-level and N-gram models of syntax. As examples of applications using this kind of knowledge, different types of part-of-speech taggers using local syntactic information were discussed. The corresponding lab exercise, spanning four lab classes, aimed at building N-gram models for use in such a system.

The last lecture of the first part of the course addressed lexical semantics with a quick glance at word sense ambiguation and information retrieval.

## 6 Applications and higher level processing

The second part of the course started with an overview lecture on natural language processing systems and finished off by a final feedback lecture, in which the course and the exam were summarised and students could give overall feedback on the total course contents and requirements.

The overview lecture addressed the topic of what makes up present-day language processing systems, using the metaphor of Douglas Adams' Babel fish (Adams, 1979): "What components do we need to build a language processing system performing the tasks of the Babel fish?" — to translate unrestricted speech in one language to another language — with Gambäck (1999) as additional reading material.

In all, the second course part consisted of nine regular lectures, two laboratory exercises, and the final evaluation lecture.

### 6.1 Machine Translation

The first main application area introduced was Machine Translation (MT). The instruction consisted of two 3-hour lectures during which the following subjects were presented: definitions and history of machine translation; different types of MT systems; paradigms of functional MT systems and translation memories today; problems, terminology, dictionaries for MT; other kinds of translation aids; a brief overview of the MT market; MT users, evaluation, and application of MT systems in real life. Parts of Arnold et al. (1994) complemented the course book.

There was no obligatory assignment in this part of the course, but the students were able to try out and experiment with online machine translation systems. Since there is no MT system for Amharic, they used their knowledge of other languages (German, French, English, Italian, etc.) to experience the use of automatic translation tools.

### 6.2 Syntax and parsing

Three lectures and one laboratory exercise were devoted to parsing and the representation of syntax, and to some present-day syntactic theories. After introducing basic context-free grammars, Dependency Grammar was taken as an example of a theory underlying many current shallow processing systems. Definite Clause Grammar, feature structures, the

concept of unification, and subcategorisation were discussed when moving on to more deeper-level, unification-based grammars.

In order to give the students an understanding of the parsing problem, both processing of artificial and natural languages was discussed, as well as human language processing, in the view of Kimball (1973). Several types of parsers were introduced, with increasing complexity: top-down and bottom-up parsing; parsing with well-formed substring tables and charts; head-first parsing and LR parsing.

### 6.3 Semantics and discourse

Computational semantics and pragmatics were covered in two lectures. The first lecture introduced the basic tools used in current approaches to semantic processing, such as lexicalisation, compositionality and syntax-driven semantic analysis, together with different ways of representing meaning: first-order logic, model-based and lambda-based semantics. Important sources of semantic ambiguity (quantifiers, for example) were discussed together with the solutions allowed by using underspecified semantic representations.

The second lecture continued the semantic representation thread by moving on to how a complete discourse may be displayed in a DRS, a Discourse Representation Structure, and how this may be used to solve problems like reference resolution. Dialogue and user modelling were introduced, covering several current conversational systems, with Zue and Glass (2000) and Wilks and Catizone (2000) as extra reading material.

### 6.4 Speech technology

The final lecture before the exam was the only one devoted to speech technology and spoken language translation systems. Some problems in current spoken dialogue systems were discussed, while text-to-speech synthesis and multimodal synthesis were just briefly touched upon. The bulk of the lecture concerned automatic speech recognition: the parts and architectures of state-of-the-art speech recognition systems, Bayes' rule, acoustic modeling, language modeling, and search strategies, such as Viterbi and A-star were introduced, as well as attempts to build recognition systems based on hybrids between Hidden Markov Models and Artificial Neural Networks.

## 7 Laboratory Exercises

Even though we knew before the course that the students' actual programming skills were not extensive, we firmly believe that the best way to learn Computational Linguistics is by hands-on experience. Thus a substantial part of the course was devoted to a set of laboratory exercises, which made up almost half of the overall grade on the course.

Each exercise was designed so that there was an (almost obligatory) short introductory lecture on the topic and the requirements of the exercise, followed by several opportunities for the students to work on the exercise in the computer lab under supervision from the lecturer. To pass, the students both had to show a working system solving the set problem and hand in a written solution/explanation. Students were allowed to work together on solving the problem, while the textual part had to be handed in by each student individually, for grading purposes.

### 7.1 Labs 1–3: Word level processing

The laboratory exercises during the first half of the course were intended to give the students hands-on experience of simple language processing using standard UNIX tools and simple Perl scripts. The platform was cygwin,<sup>1</sup> a freeware UNIX-like environment for Windows. The first two labs focused on regular expressions and the exercises included searching using 'grep', simple text preprocessing using 'sed', and building a (rather simplistic) model of Amharic noun morphology using regular expressions in (template) Perl scripts. The third lab exercise was devoted to the construction of probabilistic N-gram data from text corpora. Again standard UNIX tools were used.

Due to the students' lack of experience with this type of computer processing, more time than expected was spent on acquainting them with the UNIX environment during the first lab exercises.

### 7.2 Labs 4–5: Higher level processing

The practical exercises during the second half of the course consisted of a demo and trial of on-line machine translation systems, and two obligatory assignments, on grammars and parsing and on semantics and discourse, respectively. Both these exercises

<sup>1</sup>[www.cygwin.com](http://www.cygwin.com)

consisted of two parts and were carried out in the (freeware) SWI-Prolog framework.<sup>2</sup>

In the first part of the fourth lab exercise, the students were to familiarise themselves with basic grammars by trying out and testing parsing with a small context-free grammar. The assignments then consisted in extending this grammar both to add coverage and to restrict it (to stop “leakage”). The second part of the lab was related to parsing. The students received parsers encoding several different strategies: top-down, bottom-up, well-formed substring tables, head parsing, and link parsing (a link parser improves a bottom-up parser in a similar way as a WFST parser improves a top-down parser, by saving partial parses). The assignments included creating a test corpus for the parsers, running the parsers on the corpus, and trying to determine which of the parsers gave the best performance (and why).

The assignments of the fifth lab were on lambda-based semantics and the problems arising in a grammar when considering left-recursion and ambiguity. The lab also had a pure demo part where the students tried out Johan Bos’ “Discourse Oriented Representation and Inference System”, DORIS.<sup>3</sup>

## 8 Course Evaluation and Grading

The students were encouraged from the beginning to interact with the lecturers and to give feedback on teaching and evaluation issues. With the aim of coming up with the best possible assessment strategy — in line with suggestions in work reviewed by Elwood and Klenowski (2002), three meetings with the students took place at the beginning, the middle, and end of the course. In these meetings, students and lecturers together discussed the assessment criteria, the form of the exam, the percentage of the grade that each part of the exam would bear, and some examples of possible questions.

This effort to better reflect the objectives of the course resulted in the following form of evaluation: the five exercises of the previous section were given, with the first one carrying 5% of the total course grade, the other four 10% each, and an additional written exam (consisting of thirteen questions from the whole curriculum taught) 55%.

---

<sup>2</sup>[www.swi-prolog.org](http://www.swi-prolog.org)

<sup>3</sup>[www.cogsci.ed.ac.uk/~jbos/doris](http://www.cogsci.ed.ac.uk/~jbos/doris)

While correcting the exams, the lecturers tried to bear in mind that this was the first acquaintance of the students with NLP. Given the restrictions on the course, the results were quite positive, as none of the students taking the exam failed the course. After the marking of the exams an assessment meeting with all the students and the lecturers was held, during which each question of the exam was explained together with the right answer. The evaluation of the group did not present particular problems. For grading, the American system was used according to the standards of Addis Ababa University (i.e., with the grades ‘A+’, ‘A’, ..., ‘F’).

## 9 Results

Except for the contents of the course, the main innovation for the Information Science students was that the bulk of the course reading list and relevant materials were available online. The students were able to access the materials according to their own needs — in terms of time schedule — and download and print it without having to go to the library to copy books and papers.

Another feature of the on-line availability was that after the end of the course and as the teaching team left the country, the supervision of the students’ theses was carried out exclusively through the Internet by e-mail and chat. The final papers with the signatures of the supervisors were even sent electronically to the department. The main difficulty that had to be overcome concerned the actual writing of the theses; the students were not very experienced in producing academic text and required some distance training, through comments and suggestions, on the subject.

The main results of the course were that, based strictly on the course aims, students were successfully familiarised with the notion of NLP. This also led to eight students choosing to write their Master theses on speech and language issues: two on speech technology, on text-to-speech synthesis for Tigrinya and on speech recognition for Amharic; three on Amharic information access, on information filtering, on information retrieval and text categorisation, and on automatic text summarisation; one on customisation of a prototype English-to-Amharic transfer-based machine translation system; one on predictive SMS (Short Message Service) text

input for Amharic; and one on Amharic part-of-speech tagging. Most of these were supervised from Stockholm by the NLP course teaching team, with support from the teaching staff in Addis Ababa.

As a short-term effect, several scientific papers were generated by the Master theses efforts. As a more lasting effect, a previously fairly unknown field was not only tapped, but also triggered the students' interest for further research. Another important result was the strengthening of the connections between Ethiopian and Swedish academia, with ongoing collaboration and supervision, also of students from later batches. Still, the most important long-term effect may have been indirect: triggered by the success of the course, the Addis Ababa Faculty of Informatics in the spring of 2005 decided to establish a professorship in Natural Language Processing.

## 10 Acknowledgments

Thanks to the staff and students at the Department of Information Science, Addis Ababa University, in particular Gashaw Kebede, Kinfe Tadesse, Saba Amsalu, and Mesfin Getachew; and to Lars Asker and Atelach Alemu at Stockholm University.

This NLP course was funded by the Faculty of Informatics at Addis Ababa University and the ICT support program of SAREC, the Department for Research Cooperation at Sida, the Swedish International Development Cooperation Agency.

## References

- Douglas Adams. 1979. *The Hitch-Hiker's Guide to the Galaxy*. Pan Books, London, England.
- Atelach Alemu, Lars Asker, and Mesfin Getachew. 2003. Natural language processing for Amharic: Overview and suggestions for a way forward. In *Proceedings of the 10th Conference on Traitement Automatique des Langues Naturelles*, volume 2, pages 173–182, Bats-sur-Mer, France, June.
- Douglas Arnold, Lorna Balkan, Siety Meijer, R. Lee Humphreys, and Louisa Sadler. 1994. *Machine Translation: An Introductory Guide*. Blackwells-NCC, London, England.
- Thomas Bloor and Wondwosen Tamrat. 1996. Issues in Ethiopian language policy and education. *Journal of Multilingual and Multicultural Development*, 17(5):321–337.
- Thomas Bloor. 1995. The Ethiopic writing system: a profile. *Journal of the Simplified Spelling Society*, 19:30–36.
- Jannette Elwood and Val Klenowski. 2002. Creating communities of shared practice: the challenges of assessment use in learning and teaching. *Assessment & Evaluation in Higher Education*, 27(3):243–256.
- Samuel Eyassu and Björn Gambäck. 2005. Classifying Amharic news text using Self-Organizing Maps. In *ACL 2005 Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, Michigan, June. ACL.
- Jane Furzey. 1996. Empowering socio-economic development in Africa utilizing information technology. A country study for the United Nations Economic Commission for Africa (UNECA), African Studies Center, University of Pennsylvania.
- Björn Gambäck. 1999. Human language technology: The Babel fish. Technical Report T99-09, SICS, Stockholm, Sweden, November.
- Gerald Gazdar and Chris Mellish. 1989. *Natural Language Processing in Prolog*. Addison-Wesley, Wokingham, England.
- Raymond G. Gordon, Jr, editor. 2005. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 15 edition.
- Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey.
- John Kimball. 1973. Seven principles of surface structure parsing in natural languages. *Cognition*, 2(1):15–47.
- Fernando C. N. Pereira and Stuart M. Shieber. 1987. *Prolog and Natural Language Analysis*. Number 10 in Lecture Notes. CSLI, Stanford, California.
- Yorick Wilks and Roberta Catizone. 2000. Human-computer conversation. In *Encyclopedia of Microcomputers*. Dekker, New York, New York.
- Stephen Wurm, editor. 2001. *Atlas of the World's Languages in Danger of Disappearing*. The United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris, France, 2 edition.
- Victor Zue and James Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180.