Slovene Studies 12/1 (1990) 87-91

DOCUMENTATION
COMPUTATIONAL LINGUISTICS IN SLOVENIA: A SHORT OVERVIEW

Peter Tancig

History

The development of various computational linguistics activities in Slovenia has followed a pattern that is already well known elsewhere. *Engineering-driven work* (statistical analyses, speech recognition, artificial intelligence and natural language understanding) was being pursued from the sixties and seventies on at various technical/mathematical teaching and research departments, while *humanities-driven work* (concordances, lexicography) at different language/literature departments is of a more recent nature.

Several factors were involved in the (relative) lag in our catching up with the modern state-of-the-art of computational linguistics: the descriptive nature of domestic linguistics; insufficient funding in humanities; and the absence of a pivotal inter- and multi-disciplinary research group.

The main sites of engineering-driven work have been:

- the Faculty of Electrical Engineering in Ljubljana
- the Department of Computer Science and Informatics at the Josef Stefan Institute in Ljubljana
- the Technical Faculty in Maribor

The main sites of humanities-driven work have been:

- the Philosophical Faculty in Ljubljana
- the Institute for Slovene Language at the Slovene Academy of Sciences and Arts in Ljubljana

The Laboratory for Natural Language Understanding at JSI

The Department of Computer Science and Informatics at the Josef Stefan Institute in Ljubljana—the central Slovenian research institute in natural and technical sciences—started research in artificial intelligence at the beginning of the seventies. Some of the researchers became interested in various aspects of computational linguistics and started to work in this direction.

Getting acquainted with the modern state-of-the-art and establishing contacts with individuals and institutions abroad was, aside from shorter visits and contacts, aided especially by two events: (1) the participation of Peter Tancig at the IV International Summer School on Computational and Mathematical Linguistics in Pisa in 1977; and (2) the Fulbright Scholarship for Peter Tancig as a visiting scientist at The Laboratory for Artificial Intelligence at M.I.T., Cambridge, Mass., in the years 1980-82.

Several research projects, undergraduate and graduate theses, and co-operation with linguists were undertaken, and this developmental trend was eventually formalized by establishing the Laboratory for NLU at the end of 1986.

Research Staff

The Laboratory for NLU currently has a staff of 10, engaged full-time (for the most part) and part-time. Their professional background is mainly technical (in computer science,

mathematics), but as some have different formal education (in linguistics, psychology); this fact enables a truly inter- and multi-disciplinary approach to our work and research.

In addition to the regular members of the Laboratory, there are external personnel who are involved with ongoing research, ranging from undergraduate and graduate students to senior researchers and teachers from several departments of Ljubljana University.

Research Areas

The research staff of the Laboratory for NLU is engaged in various areas of natural language processing (NLP) with special emphasis on the specifics of the Slovene language, while dealing also with other languages (Serbo-Croat, Japanese, English). Two (standard) main streams of work are pursued:

(1) the formal modelling of various linguistic levels (phonetics, phonology, morphology, morphophonology, morphographemics, syntax, semantics and pragmatics)
(2) software and hardware tools for processing and integrating formal models of various linguistic levels into several experimental and application NLP systems

Several projects, of various degrees of sophistication and ambition, are being carried out in a number of areas, such as the following: natural (type-written) language interfacing to a DBMS, lexicography, processing text corpora for sociology and communication research, translation, and documentation. More attention has been paid lately to projects of speech analysis/synthesis, which should yield many interesting applications.

Due to the small size of the Slovene nation, which maintains all essential cultural and educational institutions, the development of several competent (and possibly competing) research groups in various fields of computational linguistics is unlikely. Thus, the Laboratory for NLU, though established and oriented towards engineering-driven research, also devotes part of its activities to humanities-driven research.

The Laboratory for NLU is also involved in various NLP research and organization activities in cooperation with other Yugoslav groups and centers for NLP, and has numerous contacts with similar research groups abroad.

Completed and Ongoing Projects

- Question-answering front-end system to a DBMS, using pattern-matching techniques
- Question-answering front-end system to a DBMS, using ATN parsing based on lexical-functional grammar and PROLOG for semantics
- Morphological analysis/synthesis based on the Koskeniemi 2-level model
- Experimental translation system, using a transfer architecture
- Phonetic analysis for formant synthesis
- Graphical development environment for formant synthesizer
- Labelled data base of spoken Slovene
- Text-to-speech system with growing complexity (letter-to-sound rules, prosody based on higher linguistic structures)
- Retrieval system of syntax structures based on (Daneu') surface syntax descriptions
- Retrieval system of syntax/semantic structures based on deep syntax descriptions (verb frames)
- Active Japanese-Slovene dictionary
- System for the analysis of large text corpora based on DBMS techniques

Yugoslav COLING Conferences

The first Yugoslav professional event which brought together people from various fields occurred in December 1977, when the Institute for Language and Literature in Sarajevo hosted a conference on Computer Processing of Linguistic Data. Twenty-one papers were presented to 65 domestic participants, with the main emphasis on computer-aided language and literature studies. The proceedings were published after the conference.

Similar Yugoslav-wide meetings were not held until October 1982 when the researchers from the Josef Stefan Institute organized a sequel in what later became a series of triennual Yugoslav conferences (Computer Processing of Linguistic/Language Data—CPLD) on NLP—covering the whole span of world COLING conferences. The most recent one (CPLD-4) was held in October, 1988. The proceedings of these conferences are the major source of information on the state-of-the-art in Yugoslavia; we try to convey the world state-of-the-art by careful selection of foreign invited speakers. In brief, the statistics of CPLD conferences are as follows:

	Yugoslav participants	Yugoslav papers	Invited foreign talks	Seminars
CPLD-2	80	39	6	0
CPLD-3	110	60	5	3
CPLD-4	124	54	6	6

Papers at the CPLD conferences can be roughly divided into the following categories:

	CPLD-2	CPLD-3	CPLD-4
statistical models	1	9	3
formal (logical) structures	14	4	6
morphological, syntactic and semantic analysis	2	8	6
software	13	3	2
computer (supported) translation	–	–	2
voice and speech recognition, analysis and synthesis of speech	1	18	15
lexicology and lexicography	7	6	6
terminological and other textual data bases	3	7	7
linguistics and literature research	1	3	2
office automation	–	–	2
other	1	–	3

Interdisciplinary Seminar on Computational Linguistics

The Laboratory has just started a regular seminar on various aspects of formal modelling and computer processing of natural language with the goal of bringing together and educating people from various disciplines and institutions. The seminar has two aspects (parts):

- longer (tutorial) presentations of various basic themes
- shorter reports on ongoing research and achieved results

Equipment

- VAX Venus at J. Stefan Institute, shared with other departments
- VAX Venus shared with other users at Ljubljana University
- several PC-ATs and compatible machines
- PS/2 model 80-171
- several microVAXes shared with other laboratories at the Department of Computer Science
- hardware and sortware for speech processing
- Software: Prolog, LISP, Pascal C, SmallTalk

University Courses

Undergraduate and graduate lectures and courses on computational linguistics are being organized (offered), albeit a bit slowly, at several university institutions in Ljubljana:

- Faculty of Electrical Engineering
- Philosophical Faculty
- Faculty for Sociology and Journalism

Other Active Institutions

Awareness of the crucial importance of various applications of more advanced natural language processing—both in written and spoken form—is gradually increasing at various places in the research community, as well as in industry. Some institutions with strong interests and beginning/ongoing projects in the field are the following:

- Faculty for Electrical Engineering and Computer Science in Ljubljana (speech analysis, speech recognition)
- Technical Faculty in Maribor (speech analysis, speech recognition)
- Slovene Academy of Sciences (lexicology and lexicography, concordances, desktop publishing)
- Board of Public Education (analyses of language of textbooks)
- Faculty for Sociology and Journalism (analyses of text corpora)
- Faculty of Medicine (dictionary of medical terminology)
- Publishing house Mladinska knjiga (encyclopedia)
- ISKRA, a major Slovene corporation in electronics (speech technology)
- GORENJE, a major Slovene producer of domestic electric and electronic appliances (speech technology)
- Radio-Television of Ljubljana (speech technology)

A European Center for Language and Speech Technologies in Slovenia

An idea—as yet very tentative—of establishing a firmer framework of international cooperation in the field is taking hold and is being discussed with various individuals/institutions in our research community and abroad, respectively. One possibility would involve UNESCO.

An international center for language and speech technologies could be set up in Ljubljana (the capital of the Republic of Slovenia with good international road connections (70 km to the Italian and Austrian borders), airport, trains, buses, communications infrastructure), or at Bled (the world famous resort, 50 km from Ljubljana). The Ljubljana site can

be established within a year as the J. Stefan Institute is just in a process of building expansion, while the Bled site still remains a viable alternative.

The European center bespeaks the (still) bridging role of Yugoslavia between Eastern and Western Europe, while the UNESCO center can be justified (and supported) by the role of Yugoslavia in the non-aligned movement.

Major Goals, Functions, Activities, Possibilities

- sponsorship of various professional associations (ESCA, ACL, ECCAI), inter-government institutions (EEC?, EC, EFTA, UNESCO, ALPE-ADRIA) and commercial companies
- clearing house (books, journals, proceedings, reports, news on various centers/activities)
- stimulating research environment for visiting individuals and small groups of researchers
- bringing together knowledge and approaches of speech processing, computational linguistics and artificial intelligence communities
- professional permanent in-house research and administrative staff
- material and administrative support (see below)
- organization of meetings (workshops, conferences, seminars, tutorials)
- evaluation and testing of research achievements and commercial products
- experimentation with commercial products
- publishing activities (proceedings/materials of meetings)
- various participating schemes for researchers from (non)-sponsoring institutions/companies

Facilities

- library
- study rooms, lecture hall(s), laboratories
- speech processing equipment
- computers (from PCs via minis to mainframes)
- links to (West) European computer networks
- software and hardware "library" of research achievements and commercial products
- permanent exhibit of commercial products
- administrative support (general secretariat and office services)

J. Stefan Institute, Ljubljana