STUDIA INFORMATICA
Nr 1/2(7) Systemy i technologie informacyjne 2006

Tradition and New Challenges for the HLT Community

Zygmunt Vetulani[1]
[1] Department of Computer Linguistics and Artificial Intelligence,
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University in Poznań, ul. Umultowska 87,
61-712 Poznań, Poland.
vetulani@amu.edu.pl; http://www.amu.edu.pl/~vetulani

Abstract. The domain of Human Language Technologies is a fascinating and challenging area of research and development. We introduce the reader into this domain, present its tradition and recent challenges.

Keywords. Human Language Technologies, Computer Linguistics, Information Society

1 Tradition

Let us start with the following working definition:

Human Language Technologies are technologies based on natural language data processing.

Human Language Technologies[1] emerged in the second half of the 20th century at the intersection of a few disciplines, the two most important among them being Computer Science and Linguistics. Let us notice that these two domains have always affected each other.

As a well identified discipline, Human Language Technologies challenge both computer science and linguistics:

- HLTs pose challenge to Computer Science forcing the latter to focus on non-numerical data and linguistic algorithms, as well as giving a new, practical dimension to the NL-oriented AI research.
- HLTs also pose a challenge to Linguistics, which must adapt its methods to the precision level, necessary for implementing language processing algorithms. Under the pressure of HLTs, linguistics has aligned in many respects to natural

[1] The term Human Language Technologies (HLT) stands for the name of the Information Society Technologies (IST) thematic programme in the Fifths Framework Programme (1998-2002). Here we will use this term in the broader, analytic sense.

sciences based on observation of empirical data (corpora studies) and scientific experiments.

We distinguish two periods in the history of the Human Language Technologies. The first one, which may be considered *classical*, and, which determines the tradition of the discipline, ends in mid 80-ties. The second one continues until now. During this first period, the term Human Language Technologies was not in use. Problems typical of this domain used to be identified as belonging to cybernetics, artificial intelligence and finally *computational linguistics.*

As working definition of computational linguistics we may take the following:

Computational Linguistics is a discipline aiming at computer simulation of human verbal communicational[2] competence[3].

1.1 Beginnings of Human Language Technologies: Computational Linguistics

Computational linguistics since the very beginning has been marked by the ambitious project of machine translation. Indeed, machine translation remained the main "human language technology" for a long time.

As early as in 1946 A.D.Booth (Richens & Booth 1955), the head of the laboratory of electronic computing in London started his first works on the automatic dictionary and advocated a large-scale research on machine translation. He convinced of this idea Warren Weaver, a cryptologue and vice-president of the Rockefeller Foundation. His famous "Memorandum" of July 15, 1949 is considered essential for mobilisation of important financial means for MT research, first of all in the USA[4].

Let us note here that computational linguistics has a long prehistory. The letter of René Descartes to father Mersenne of October 26, 1629 is considered a herald of machine translation. Descartes postulated in this letter a numerical dictionary as support for "mechanical" translation between languages (Mounin 1964).

Also, L. Couturat and L. Leau (1903) mention the lost paper by W. Rieger (XVII century) entitled "Zifferengrammatik, welche mit Hilfe der Wörterbtcher ein mechanisches Ubersetzen aus einer Sprache in alle andere ermöglicht" ("Code-grammar which, with the help of dictionaries, enables mechanical translation from one language into all others")[5]. In the 30s Turing and Smirnov-Trojanskij[6] wrote

[2] The notion of communicative competence was first identified by Halliday (cf. (Halliday 1970)) and http://www.ne.jp/asahi/kurazumi/peon/ccread.htm). The definition formulated by Brown is: "Communicative competence (...) is that aspect of our competence that enables us to convey and interpret messages and to negotiate meanings within specific contexts" (Brown 1987,1994).

[3] We will make abstraction of the very moment of the first use of the term c*omputational* linguistics and apply it for the whole period we are interested in.

[4] The "Memorandum" is reproduced in (Locke and Booth 1955).

[5] Cf. the findings of Couturat and Leau (Couturat and Leau 1903), cf. also Hutchins (http://ourworld.compuserve.com/homepages/WJHutchins).

about the idea of mechanical translation. Pioneering works of the latter author remained unknown until 1951.

After the first spectacular achievements in the 50s in the USA (Georgetown, cf. (Sheridan 1955)) and in the Soviet Union (Moscow, cf. (Panov 1956)), it appeared that obtaining high quality machine translation of unrestricted texts (or speech) is one of the hardest problems of applied computer science (AI). 50 years later we are still far from the final goal.

Machine translation was the first but not the only one "strategic objective" in computational linguistics in the *classical* period. Computational Linguistics was stimulated by other challenges inspired by cybernetics, artificial intelligence, robotics and even the science fiction literature. In particular, the vision of humanoid "intelligent" robots was at the origin of very dynamic research on man-machine communication. This field appeared even more complex than machine translation, defined as language-to-language transformation that can be mathematically described. Man-machine communication involves, in particular, speech recognition as well as computer modelling of understanding and reasoning.

Typical problems of computational linguistics are:

- machine translation,
- natural language communication with robots,
- natural language access to systems for storing and processing information,
- natural language access to interactive aid systems,
- automatic generation of technical documentation,
- text processing (text generation, summarisation, information retrieval and extraction, error detection and correction).

In what follows we provide a few examples of projects that have considerably influenced research and technologies in Human Language Technologies in the first, classical, period:

- BASEBALL (B. Green, A. Wolf, C. Chomsky, K. Laughery, University of California, 1961) one of the first question-answering systems (knowledge representation is based on frames, syntactic analysis on the ground of the works by Harris)[7].
- ELIZA (J. Weizenbaum, 1966) a system for conversation maintenance based on pattern-matching, aiming at the surface simulation of a dialogue and of the quality which would make the system pass the Turing test (in contrast to the common belief, the dialogue maintenance systems may have some interesting practical applications)[8].
- LUNAR (W.A.Woods, BBN, 1972) a system for consulting a database about the samples taken from the Moon by the Apollo 11 vehicle (ATN, procedural semantics)[9],

[6] Cf. (Hutchins and Lovtskij 2000).
[7] Cf. (Green, Wolf, Chomsky and Laughery 1961).
[8] Cf. (Weizenbaum 1966).
[9] Cf. (Woods 1978).

- SHRDLU (T. Winograd, MIT, 1972) a system for controlling a robot supposed to move geometrical objects (functional grammar of Halliday, procedural semantics, cognitive system written in PLANNER)[10],
- LADDER (G. Hendrix, E. Sacredoti, D. Sagalowicz, J. Slocum, SRI, 1977) a dialogue-based access system to the distributed data base (semantic grammars)[11],
- GUS (D.G. Bobrow, Kaplan, M. Kay, Norman, Tompson, T. Winograd, Xerox Palo Alto, 1977) task oriented dialogues (transition network grammar, case grammar (Ch. Fillmore), frames, application of object programming principles /procedural attachment/, frame based dialogue control)[12].
- PARRY (R.C. Parkison, K.M. Colby, W.S. Faught, Univ. California, 1977) - computer model of paranoia[13].
- TEAM, DIALOGIC (P. Martin, D. Appelt, F. Pereira, B. Grosz,... SRI, 1983) - portable system of data base access derived from the LADDER system (separation of syntax and semantics, auto adaptable to the given data base).[14]
- ELI - English Language Interpreter (Riesbeck), QUALM - Module Q/A (W. Lehnert), SAM - Script Applier Mechanism (R. Cullingford, R. Schank), 70ties, Yale; SAM processes stories read by ELI and answers user's questions (QUALM) making use of the knowledge representation mechanisms based on the memory model proposed by Schank (using *situations*, *scripts* and *episodes*)[15].
- PAM - Plans Applier Mechanism (R. Wilensky, ok. 1980), a system for reading and processing stories, uses the Schank concepts of memory organisation (the memory model organised by *turning-points*; text grammar)[16].
- HAM-ANS (W. Hahn, W. Hoeppner, K. Morik, H. Marburger and others, Hamburg, 1981-1986) a dialogue system based on an integrating approach to language processing (the syntactic, semantic and pragmatic components are not separable); hotel reservation dialogue based on user modelling; a two-layer knowledge representation (conceptual and referential knowledge)[17]. Since 1986 to 1989 continued as WISBER.
- ORBIS (A. Colmerauer, R. Kittredge, Marsylia, early 80ties) bilingual system answering English or French questions about planets and other astronomical objects, implemented entirely in Prolog II (Marseille PROLOG) in order to demonstrate the strength of this language[18].
- The Polish module ORBIS-PL[19] (implemented by Z. Vetulani in 1984) was the starting point to the work on much more efficient understanding systems for

[10] Cf. (Winograd 1973).
[11] Cf. (Hendrix, Sacredoti, Sagalowicz and Slocum 1978).
[12] Cf. (Bobrow 1977).
[13] Cf. (Parkison, Colby and Faught 1977).
[14] Cf. (Martin, Appelt and Pereira 1983).
[15] Cf. (Cullingford 1981).
[16] Cf. (Wilensky 1977).
[17] Cf. (Hoeppner, Morik and Marburger 1986).
[18] Cf. (Colmerauer and Kittredge 1982).
[19] Cf. (Vetulani 1988).

Polish in form of various versions of the POLINT system (in development until now)[20].

All these systems were prototypes with no direct practical follow-up.

1.2 Methodological challenge

The basic methodological challenge of the first period was to establish methodological basis for the new discipline. Let us quote here two opinions, from two different epochs of the classical period.

- S. Ceccato (one of the machine translation pioneers, 1956) postulated "research on the nature of thought (...) with the objective to construct artefacts able to perform some of our mental operations and give them a mental expression ". [21]
- R. Schank (one of pioneers of Cognitive Science) wrote the following in 1980 in "Language and Memory": "The theory I have been trying to build here is an attempt to account for the facts of memory to the extent that they are available (...) I do not believe that there is any other alternative available to us in building intelligent machines other than modelling people."[22]

Let us remark that the position of Schank is very clear and goes far beyond the requirements of Turing style methodology where the *Turing test* is considered as the basic intelligence measuring tool. Still, "modelling people" continues to be a weak point of this methodology because today we do not have a satisfactory knowledge about basic human mental aptitudes (recognition, logical inference, decision taking). The existing theories are speculative and vague. Also, there is a lack of experimental and observational research to give a solid basis for such a theory. This problem was identified a long time ago and motivated a number of CL researchers to try and fill the gap.

We will quote here a few examples of application of the methodology drafted above the research, typical of the early HLT.

- SRI (B. Grosz) - observations and analysis of experimental task-oriented dialogues, studies of thematic-rhematic dialogue structure in terms of attention focussing, etc.,[23]
- Hopkins University (A. Chapanis) - experimental research on correlations between language performance and information channels and modes[24],
- AUM (R. Kittredge) - research on sublanguages from the point of view of machine translation feasibility[25],

[20] Cf (Vetulani 1997) and (Vetulani 2004).
[21] "des recherches sur la nature de la pensée (...) en vue de construction d'un appareil qui puisse exécuter certaines de nos opérations mentales et leur donner une expression mentale", after (Mounin 1964).
[22] Cf. (Schank 1980).
[23] Cf. (Grosz 1977).
[24] Cf. (Chapanis 1973) and (Chapanis 1975).
[25] Cf. (Kittredge 1982).

- University of California - simulation of the man-system dialogues concerning flights (as a part of the GUS - Xerox Palo Alto project)[26],
- WISBER (Hamburg) - studies of dialogue structure[27].

The alternative solution is the "black box" methodology where the internal structure of the phenomena being modelled is considered entirely or partly unknown, and where the project designer has the Turing test as its only criterion for system validation.

2 New challenges: towards the Information Society

The methods and results of the *classical* periods have mostly not become out-of-date, have not become forgotten and are still being developed and improved. On the other hand, priorities in Human Language Technologies have changed.

In contrast to the challenges of the first, classical period, which have the character of technical achievements (intended to demonstrate what can be done), new challenges have technological character and have been triggered by practical needs. They are closely linked to geopolitical changes all over the world and to the process called globalisation.

By *globalisation* we mean breaking down borders and divisions by political, technical, economical and cultural thought. Globalisation, though already known in the past (despite the lack of today's technical measures), is a new phenomenon characterised by the unobserved (until now) flow of information and mobility of people (it is interesting to see that these aspects of globalisation are explored by both its fans and opponents). Conceived in the 90s, the idea of the Global Village seems now feasible thanks to the progress of *communication technologies*, both in the traditional sense of mobility of goods and persons, and in the sense of information transfer technologies (telecommunications, teleinformatics). Development of network technologies played an essential role. The first spectacular success in this area was that of the French MINITEL[28] system. It was launched in 1981 and positively tested in France on national scale. MINITEL permitted to implement concept of network services, starting with the famous "3615" (directory) service. The experiment was successful thanks to large access to terminals distributed for free to France Telecom customers (who until that time had not been computer users in most cases). This success resulted in universal computer education of French people, however, did not have much impact in other countries because of the arrival of much more powerful Internet and general availability of cheap personal computers.

At the same time, political changes in Europe, especially the symbolic fall of the Berlin Wall on November 9th, 1989, created in Europe new political climate favourable to enhancing European integration. One of the great integrating ideas that emerged in the 90s was the announcement by the European Commission of

[26] Cf (Bobrow et al. 1997).
[27] Cf. (Gerlach, Horacek 1989).
[28] Cf. (Rincé 1990).

a programme to transform Europe into an Information Society[29]. The objective of this programme was to find through science and technology a solution to the discrepancy between the wish to enhance free access to information (in order to increase competitiveness of economy in the global village) and the wish to maintain multicultural and multilingual aspect of Europe as a part of our precious cultural heritage.

As a big challenge of the turn of the centuries, in particular, changing the way of thinking about computational linguistics, we consider creation of the infrastructure, which will be the foundation for building a multilingual and multicultural Information Society.

What follows is the challenge to build in Europe a strong and competitive language industry able to produce the Information Society infrastructure

New definition:

By *Human Language Technologies* we mean technologies used to build informatics linguistic infrastructure for the Information Society.

This challenge may be characterised in a more abstract way, without recurring to socio-political categories. Namely, Human Language Technologies may also be seen as the technologies of interaction between a human and his technological environment. This environment is changing rapidly. Until recently, it was information empty and its components were static inactive artefacts. Now the situation is quite different. The human's technical environment, initially produced by man, has become an extension of natural environment with its own autonomy. Elements of this environment, like the Internet, seem to have their own identity, highly independent of the individuals and even organisations. This environment is saturated by information (information-rich). In this new situation, humans may wish to communicate with this environment as they do with other humans. Natural language technologies are there in order to provide this environment with language competence compatible with the human natural language competence. Providing means for such communication in the situation of dynamic evolution of the technological environment constitutes a challenge for Human Language Technologies considered as a part of Artificial Intelligence (in the broad meaning of this term).[30]

2.1 Electronic resources of Human Language Technologies

The new challenge presented above implies a new way of thinking about objectives. The postulated infrastructure has to include the technological components

[29] The IST Program (Information Society Technologies, also called User-friendly Information Society), 1998-2002, within the 5FP, with the budget of 3600 MECU.
(http://europa.eu.int/comm/research/ist/leaflets/en/intro2.html)

[30] This paragraph summarises my contribution to the "Technology for Linguistics, Linguistics for Technology" panel discussion co-hosted by Language and Technology 2005 and PLM 2005 conferences (in: (Vetulani 2005)).

derived from the existing laboratory prototypes but able to work in real situations and in real time. The necessary condition in order to meet these latter requirements is availability of necessary language resources. The concept of language resources (LR) was "invented" and promoted by the visionary pioneer of language industries Antonio Zampolli[31]. Zampolli defined this concept as meaning "written or spoken corpora, lexical data bases, grammars"[32]. It is important to say that the identification of real needs concerning operational tools (not merely prototypes) induced a methodological change in the area of linguistics consisting in abandonment of the "tendency (dominating in linguistics in the 70s and in the early 80s) to test research hypothesis on the basis of a small number of (allegedly) critical importance data." (Zampolli, ibid.)

The new approach whose pioneers in Europe were the Italian researcher Antonio Zampolli (Calzolari 2005) and the French scientist Maurice Gross (Laporte 2005) contributed to the rapprochement between the methodology of linguistics and the methodology of natural sciences. It postulates constructing systems with some language competence (as translating systems, summarising systems, correctors, speech analysers) which work in real time and in real world. Such systems should be subjects of investigations using observation and scientific experiments. These postulates of constructing language resources (but also standards, formalisms, tools exploring these resources and tools to obtain them) were realised in many projects, first of them being inspired by the famous Grosseto Workshop (On Automating the Lexicon) organised by A. Zampolli, N. Calzolari and D. Walker in the year 1986[33]. Let us mention some of them that have impact on language technologies.

- Acquilex I and II - 1989-1995 "explore the utility of constructing a multilingual lexical knowledge base from machine-readable versions of conventional dictionaries"
  (cf. http://www.cl.cam.ac.uk/Research/NL/acquilex/acqhome.html).
- ESPRIT MULTILEX 1990-1993: research and development project aiming at providing specifications of standards for multilingual lexicons (cf. http://www.ilc.cnr.it/EAGLES96/edintro/node11.html).
- EUREKA GENELEX (1990-1994) program which aimed at developing a general-purpose dictionary format independent of theories and applications[34]. It was extended by the PECO/COPERNICUS project CENTRAL EUROPEAN GENELEX MODEL (CEGLEX, 1995-1996)[35]
  (http://www.kc.t.u-tokyo.ac.jp/NLP_Portal/initiative-e.html
  http://dbs.cordis.lu/cordis-cgi
  /srchidadb?ACTION=D&SESSION=199552002-3-6&TBL=EN_PROJ
  &RCN=EP_RCN:29812

[31] Cf. (Calzolari 2005).
[32] Cf. (Zampolli 1996).
[33] Cf. (Walker, Zampolli and Calzolari 1994).
[34] Cf. (Antoni-Lay, Francopoulo and Zaysser 1994).
[35] Cf. (Vetulani 2000).

http://www.amu.edu.pl/~zlisi/projects/ceglex/index.en.html).

- MULTEXT (Multilingual Text Tools and Corpora) was intended to contribute to the development of generally usable software tools to manipulate and analyse multi-lingual text and speech, and to annotate multi-lingual text and speech corpora with structural and linguistic markup (cf. http://www.isca-speech.org/archive/ssw2/ssw2_077.html).
- RELATOR (1994-1995) was "a European-wide consortium of researchers who, with the support of the European Commission, striving to establish a European repository of linguistic resources" (cf. http://www.dfki.de/lt/projects/relator.html). RELATOR resulted in the ELRA association.
- TEI "Initially launched in 1987, TEI is an international and interdisciplinary standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent." (http://xml.coverpages.org/tei.html and http://www.tei-c.org/)
- EAGLES/ISLE (EAGLES - European Advisory Group on Language Engineering Standards, 1993-1999; ISLE - International Standards for Language Engineering, European-US joint project, 2000-2002).
- LE-PAROLE project (1996-1998) aimed to "offer a large-scale harmonised set of "core" corpora and lexica for all European Union languages". (http://www.elda.org/catalogue/en/text/doc/parole.html).
- SIMPLE project (1998-2000) "The goal of SIMPLE project is to add semantic information, selected for its relevance for LE applications, to the set of harmonised multifunctional lexica built for 12 European languages by the PAROLE consortium." (http://www.ub.es/gilcub/SIMPLE/simple.html, http://www.ilsp.gr/simple_eng.html)
- WORDNET (a lexical database for English where words are organised into synonym classes and hierarchies)[36] and EuroWordNet (multilingual database with wordnets for various European languages, EU funded project inspired by WORDNET)[37].

2.2 Building language industries in Europe

The appeal by European Commission to build an Information Society puts emphasis on creating basis of language industries. An important part of the necessary effort is creation of language resources necessary to verify theoretical results (e.g. language corpora) but first of all to design the systems involving natural language processing (lexica, thesauri, grammars) and to validate such systems.

Building the language industry has become a priority in the technologically leading countries and especially in the USA, Japan, some EU countries but also in

[36] http://wordnet.princeton.org
[37] http://www.illc.uva.nl/EuroWordNet/

China (our knowledge about the involvement of the latter country is limited). In this talk, we will focus on the European efforts within the confines of the rivalry with the USA and Japan.

In the beginnings of language industries in Europe an important stimulating role was played by the translational initiatives. Among one of the first such initiatives we have to mention EUREKA programme (European Research Co-ordination Agency) meant as an instrument to enhance competitiveness of Europe in this field through the enhancement of market driven research. In the 10-year period of 1986-1995, this programme was realised by over 1000 companies organised into the consortia covering 22 countries and with a budget exceeding 10 billion ECU. Among ca 30 information technology projects at least 4 were specifically oriented towards the language engineering needs. (E.g. EUREKA-GENELEX with the budget of 37,7 MECU, EUREKA-EUROLANG with the budget of 69MECU, according the Language Industries Atlas)[38].

Parallel language technology projects were funded by successive CE Framework Programmes (FP). In 1984, the European Commission launched the ESPRIT programme (European Strategic Programme for Research and Development in Information Technology) within the first FP with the following objectives: (1) "to promote the co-operation between industrials, research centres and universities in the field of information technologies, (2) to accelerate the development of basic European technology in order to increase international competitiveness and (3) to achieve international recognition for the technical standards for the IT market." (after the Language Industries Atlas). In the years 1984-1994, the ESPRIT programme supported ca. 70 language technology projects with ca. 200 MECU.

Within the 3rd FP (1990-1994), under the Linguistic Research and Engineering programme (LRE), the following 3 areas were prioritised (with the emphasis on building theoretical foundations of language technologies):

- "General research, to tackle the many remaining research problems and foster progress to more sophisticated language understanding technologies,
- Common resources, tasks and methods to build over time a comprehensive infrastructure,
- Pilot applications, to demonstrate the integration of language engineering technologies and components within information and communication systems."[39]

Within the 4th Framework Programme the focus was shifted from theory to practical commercially exploitable applications. Within the "Telematics" thematic programme the very precise objectives of building written, spoken and terminological resources were defined e.g. concerning written resources, for the following were priority tasks[40]:

[38] Cf. (Hearn and Button 1994).
[39] Cf. *Language and Technology. From the Tower of Babel to the Global Village.* Brochre published by the Office for Official Publications of the European Commission, Luxembourg, 1996 (ISBN 92-827-6974-7) (page 19).
[40] According to Zampolli (Zampolli 1996).

- creation of "monolingual dictionaries containing min. 50.000 lexemes each, for at least the 11 EC official languages, harmonised in the way easing exchangeability, common efficiency and useful for building monolingual interfaces in the future,
- creation of "text corpora for the languages mentioned above containing min. 50.000.000 words each, as a basis for dictionary creation and maintenance; if possible parallel multilingual text corpora,
- creation of "integrated tools for linguistic coding, analysis, search and evaluation".

The ventures inspired by the European institutions are usually provided with substantial funding (cf. EUREKA, above). Besides money, an essential organisational effort was made, which resulted in research institutions, academic curricula, societies and large-scale conferences. Let us provide some examples of language technology specialised institutes:

- Instituto di Linguistica Computazionale, founded by Antonio Zampolli in Pisa as one of the first institutes of that kind in the world,
- Centre for Language Technologies (Center for Sprogteknologi), established in 1991 in Copenhagen (and affiliated to the Copenhagen University),
- Institute for Language and Speech Processing (ILSP) established in 1991 in Athens under the auspices of Hellenic General Secretariat of Research and Technology (by G. Carayannis).

The US earlier initiatives such as

- Association of Machine Translation and Computational Linguistics founded in 1962, since 1968 as Association of Computational Linguistics (http://www.aclweb.org),
- COLING (60s) - informal organisation named International Committee on Computational Linguistics having as its main objective organisation of International Conferences on Computational Linguistics (COLING) (http://www.dcs.shef.ac.uk/research/ilash/iccl/)

were followed by European language industry oriented initiatives. We list some of them below:

- In 1991 the European Association for Machine Translation was registered in Geneva (Switzerland) as a "non-profit" institution (http://www.eamt.org/),
- In 1995 the European Language Resources Association (ELRA) (http://www.elra.info/) was registered in Luxembourg (at the DGXIII inspiration); ELRA operates through its agenda for gathering and distributing of language resources ELDA (Evaluation and Language Resources Agency) (http://www.elda.org/sommaire.php) (ELRA resulted from the RELATOR project).
- "Excellence networks", such as ELSNET (European Network of Excellence in Human Language Technologies) with its head office installed in Utrecht (http://www.elsnet.org/) in 1991, were established for the integration purposes.

An essential activity of international organisations is organisation of meetings. The leading conference cycles such as the Annual Meetings of the ACL or

COLING (sometimes organised as joint events, such as the conference 21st COLING and 44th ACL Annual Meeting planned for 2006) were completed by the LREC (Language Resources and Evaluation Conference) "invented" by Zampolli in 1998. The LRECs, organised every 2 years by ELRA, has become the main conferences in the area of language resources (with over 800 participants at the Lisbon meeting in 2004). In Poland, the conference "Language and Technology: Human Language Technologies as a Challenge for Computer Science and Linguistics, April 21-23, 2005, Poznań" was very successful with 150 participants from all over the world; this initiative will be continued (http://www.ltc.amu.edu.pl).[41]

2.3 The new challenge

The information provided above is to illustrate the huge financial and organisational effort made by the EU countries and international bodies by the end of the 20th century but also to show dangers involved. A real danger results from the fact that the funding of research and development on European scale is limited to the actual priorities. These priorities change from one framework programme to another. E.g. in the 5th and 6th FPs the construction of language resources is no more an objective as such. What has become a priority is the practical application of technologies (feeding the idea of Information Society). Also in the forthcoming 7th FP the focus will change with respect to the former FP as declares the Commissioner for Science and Research Janez Potočnik: "Evidently, we cannot forget that research for research's sake is not the objective of the framework programme - we need to ensure that the results are used. (...) This is why we are placing much more emphasis on promoting knowledge transfer and the use of research results in FP7"[42]. Such a policy speeds up the progress favouring the beneficiary countries with respect to all others. This policy generated, however, negative side effects, in particular, for the new EC member states which were not covered by the 3rd and 4th FPs, and which could not afford a parallel effort financed by themselves. The EC was partially conscious of the problem and extended the awareness operations consisting in organising conferences "Language and Technology Awareness Days" to the UE candidates. (The conference "Language and Technology Awareness Days, 1995 Poznań, Poland" was organised by myself under the EC funding. It gathered together over 100 participants from Poland.) Also, some financial support (relatively modest) was provided under the programs like PECO-COPERNICUS opened to mixed EU-CEC consortia (e.g. the GRAMLEX and CEGLEX projects[43] were financed within this scheme). These measures had only very limited effect, and it is hard to consider their impact with respect to the international competition as satisfactory from the point of view of the countries concerned. The problem of a still existing (not to say growing in some areas) gap between the countries of the "old" European Unions, and the "new" member countries resulting from the lack of synchronisation between the

[41] Cf. (Vetulani 2005).
[42] In "Potočnik pushes exploitation of knowledge up the agenda", Cordis Focus, No 256, June 2005, p.18.
[43] Cf. (Vetulani 2000).

EC programs, and the needs and potential of the candidate countries (today's new members) was articulated by myself at the panel discussions of the LREC 1998 (Grenada) and LREC 2000 (Athens) meetings. I suggested more institutional effort (both financial and organisational) in order to help the countries concerned to reach the excellence level of the leading countries in particular in the domain of basic language resources.

Lack of such operations (or of the political will to operate) on European scale poses a new challenge to each country concerned (including Poland). Answering this challenge should be considered priority. Zampolli predicted the present situation already 10 years ago in his text read at the L&T'95 in Poznań[44]:

- "LRs are closely related to the national and cultural identity and play crucial infrastructural role in obtaining language industry products for a given language",
- "it is commonly understood that the existence of language industries constitutes a necessary condition for preserving language as the communication support in the contemporary information society".

Zampolli also claimed - in accordance with the EC viewpoint - that "promotion of language resources for a given language is a task for the competent national administrations", and that "language resources should be available as public domain property".

Conclusion

Building national electronic language resources as a basis for language engineering and for national language industries at a level satisfying needs of the international competitiveness and permitting construction of the global Information Society base including all national languages is the challenge first of all for the national research communities and the respective state administrations. Nevertheless, it also poses challenge for pan-national administrations covering countries aiming to accede to the global Information Society, in particular for the recently enlarged Europe.

References

1. Antoni-Lay M.-H., Francopoulo G. and Zaysser L., (1994). A generic model for reusable lexicons: The GENELEX project, *Literary and Linguistic Computing* 9(1). 47-54.
2. Bennett W.S. and Slocum J., (1985). The LRC Machine Translation System. *Computational Linguistics* 11 (2-3), 111-121.
3. Bobrow D.G., Kaplan R.M., Kay M., Norman D.A., Thompson H. and Winograd T., (1977). GUS, A Frame Driven Dialog System, *Artificial Intelligence* 8, 155-173 (reproduced in (Grosz 1986)).

[44] Cf. (Zampolli 1996).

4. Calzolari N., (2005). Antonio Zampolli, a life for Computational Linguistics. In: *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proc. of Language and Technology Conference, April 21-23, 2005, Poznań, Poland,* Vetulani Z. (ed.), Wyd. Poznańskie, p. XXIII.
5. Chapanis A., (1973). The Communication of factual Information through Various Channels, *Inform. Stor. Retr.*, Vol. 9., 215-231.
6. Chapanis A., (1975). Interactive Human Communication, *American Scientist,* March 1975, 36-42.
7. Colmerauer A., (1970). *Les systèmes Q, ou un formalisme pour analyser et synthétiser les phrases sur l'ordinateur*. Publication interne No. 43. TAUM. Université de Montreal.
8. Colmerauer A. and Kittredge R., (1982). ORBIS, *Proceedings of the 9th COLING Conference*.
9. Couturat L. and Leau L., (1903). *Histoire de la langue universelle*. Paris, France, Hachette.
10. Cullingford R., (1981). SAM. In: *Inside Computer Understanding,* Schank, R. and Reisbeck C. (eds.), Hillsdale: Lawrence Erlbaum Associates (reproduced in (Grosz 1986)).
11. Fillmore C. J., (1968). The case for case. In: *Universals in linguistic theory*, Bach E., Harms R.T. (eds.), New York: Holt, Rinehart and Winston.
12. Gazdar G., Klein E., Pullum G.K., and Sag I.A., (1985). *Generalized Phrase Structure Grammar*. Oxford: Basil Blackwell.
13. Gerlach M., Horacek H., (1989). Dialog Control in a Natural Language System. *EACL 1989*, 27-34.
14. Green B., Wolf A., Chomsky C., Laughery K., (1961). BASEBALL: An Automatic Question Answerer, *Proceedings of the Western Joint Computer Conference* 19, pp. 219-224 (reproduced in (Grosz 1986)).
15. Gross M., (1984). Lexicon-Grammar And The Syntactic Analysis Of French. *Coling 1984*: 275-282.
16. Grosz B., (1977). The Representation and Use of Focus in a System for Understanding Dialogs, *IJCAI 1977*, 67-76.
17. Grosz B.J., Sidner C.L., (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12 (3), 175-204.
18. Grosz B. et al. (eds.), (1986), *Readings in Natural Language Processing*, Morgan Kaufmann Publishers, Inc., Los Altos, California.
19. Halliday M.A.K., (1970). Language structure and language function. In: *New Horizons in Linguistics*, J. Lyons (Ed.), London: Penguin Books.
20. Hearn P. and Button D. (eds.), (1994). *Language Industries Atlas*, IOS Press, Amsterdam, Oxford, Washington, Tokyo.
21. Hendrix G., Sacredoti E., Sagalowicz D., and Slocum J., (1978). Developing a Natural Language Interface to Complex Data, *ACM Trans. on Database Sys.* 3(2), pp. 105-147 (reproduced in (Grosz 1986)).

22. Hoeppner W., Morik K., Marburger H., (1986). Talking it Over, The Natural Language Dialog System HAM-ANS. *Cooperative Interfaces to Information Systems 1986*, 189-258.
23. Hutchins J., (1997). From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947-1954. A Chronology. *Machine Translation*, Volume 12, Number 3, 195 - 252.
24. Hutchins J. and Lovtskij E, (2000). Petr Petrovich Troyanskij (1894-1950). a forgotten pioneer of machine translation. *Machine Translation* 15(3), 187-221.
25. Kaplan R.M and Bresnan J., (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In: *The Mental Representation of Grammatical Relations*, Bresnan, J. (ed.), Cambridge MA, 727-796.
26. Kay M., (1985). Parsing in Functional Unification Grammar. In: *Natural Language Parsing*, D. Dowry, L. Karttunen, and A. Zwicky (eds.), Cambridge University Press, Cambridge, England, 251-278.
27. Kay M., (1996). *Machine Translation: The Disappointing Past and Present.* Xerox Palo Alto Research Center, Palo Alto, California, USA (http://www.fortunecity.com/business/reception/19/xparc.htm).
28. Kittredge R.I., (1982). Sublanguages, *American Journal of Computational Linguistics 8 (2)*, 79-84.
29. Laporte E., (2005). *In memoriam* Maurice Gross. In: *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proc. of Language and Technology Conference, April 21-23, 2005, Poznań, Poland*, Vetulani Z. (ed.), Wyd. Poznańskie, p. XX.
30. Locke W.N. and Booth A.D., (1955). *Machine translation of languages*. MIT Press, Cambridge Mass.
31. Martin P., Appelt D., Pereira F., (1983). Transportability and Generality in a Natural-Language Interface System. In: *Proc. of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany*, Los Altos: Wiliam Kaufmann, Inc. 573-581 (reproduced in (Grosz 1986)).
32. Morik K., (1985). User Modelling, Dialog Structure and Dialog Strategy in HAM-ANS, *EACL 1985*, 268-273.
33. Mounin G. (1964). *La Machine à traduire*. La Haye: Mouton (in French)).
34. Panov D.Iv., (1956). *Automatic Translation*, Moscow: Izdatel'stvo AN SSR (in Russian).
35. Parkison R.C., Colby K. M., Faught W. S., (1977). Conversational Language Comprehension Using Integrated Pattern-Matching and Parsing, *Artificial Intelligence* 9, 111-134 (reproduced in (Grosz 1986)).
36. Richens R.H. and Booth A.D., (1955). Some methods of mechanized translation. In: *Machine translation of languages: fourteen essays*, Locke W.N. and Booth A.D. (eds.), Cambridge, Mass.: The Technology Press of the Massachusetts Institute of Technology, 24-46.

37. Rincé J.-Y., (1990). *Le MINITEL*, Que sais-je?, 2539, Presse Universitaire de France, Paris.
38. Schank R.C., (1980). Language and memory, *Cognitive Science*, 4, 243-284.
39. Sheridan P., (1955). Research in Language Translation on the IBM Type 701, *IBM Technical Newsletter*, No. 9, IBM, New York (Oct 1955), pp. 95-104.
40. Slocum J., (1985). A machine translation bibliography (Generally restricted to currently accessible documents written in English, French, or German during the years (1973-1984), *Computational Linguistics*, Vol. 11, Numbers 2-3, April-September 1985.
41. Vetulani Z., (1988). PROLOG Implementation of an Access in Polish to a Data Base, *Studia z automatyki*, XII, PWN, 5-23.
42. Vetulani Z., (1997). A system for Computer Understanding of Texts. In: *Euphony and Logos*, R. Murawski, J. Pogonowski (eds.), Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 57, Rodopi, Amsterdam-Atlanta, 387-416.
43. Vetulani Z., (2000). Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX. In: *Second International Conference on Language Resources and Evaluation, Athens, Greece, 30.05.-2.06.2000, (Proceedings)*, Gavrilidou M. et al. (eds.), ELRA, Paris, 367-374.
44. Vetulani Z., (2004). *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej* (in Polish), Akademicka Oficyna Wydawnicza EXIT, Warszawa. (English title: Man-Machine Communication: Computer Modelling of Human Language Competence).
45. Vetulani Z. (ed.), (2005). *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proc. of Language and Technology Conference, April 21-23, 2005, Poznań, Poland*. Poznań: Wyd. Poznańskie, pp. XXVI-XXX.
46. Walker D., Zampolli A., Calzolari N. (eds.), (1994). *Automating the lexicon: research and practice in a multilingual environment*. Oxford: OUP.
47. Weizenbaum J., (1966). ELIZA - A Computer Program for the Study if Natural Language Communication Between Man a Machine, *Communications of the ACM*, 10, pp. 36-43.
48. Wilensky R., (1977). PAM - A Program That Infers Intentions. *IJCAI 1977*: 15.
49. Wilensky R., (1983). Memory and Inference. *IJCAI 1983*, 402-404.
50. Winograd T., (1972). *Understanding Natural Language*, Academic Press, New York.
51. Winograd T., (1973).: A procedural Model for Language Understanding. In: *Computer Models of Thought and Language*, R. Schank and K. Colby (Eds.), 152-186 (reproduced in (Grosz 1986)).
52. Woods W.A., (1970). Transition Network Grammars for Natural Language Analysis, *Comm. of the ACM*, 13, 10.

53. Woods W.A., (1978). Semantics and Quantification in Natural Language Question Answering, *Advances in Computers*, vol. 17, Yovits, M. (ed.), 2-64, New York: Academic Press (reproduced in (Grosz 1986)).
54. Zampolli A., (1996). Współpraca międzynarodowa w dziedzinie LR (in Polish), *Informatyka*, Nr 3, 34-37. (English title: International co-operation in the domain of Language Resources).