Emulated Language Competence and Other Selected Works in the Field of Polish Language Engineering at the Department of Computer Linguistics and Artificial Intelligence of the Adam Mickiewicz University

Presentatuion for the HLT Days, Warsaw, September 27–28, 2012

by

Zygmunt Vetulani[1], Grażyna Vetulani[2], Tomasz Obrębski[1], Jacek Marciniak[1]

from

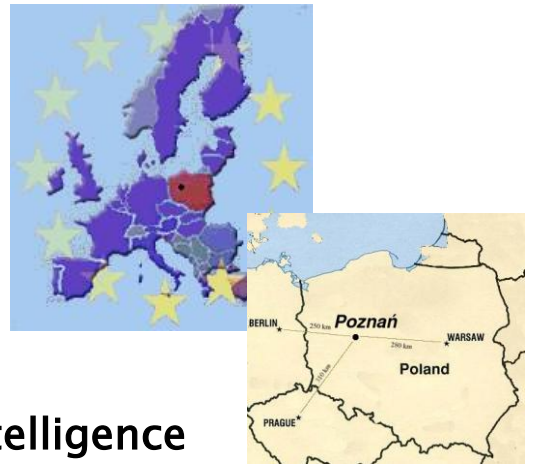
Adam Mickiewicz University in Poznań
Dept of Computer Linguistics and Artificial Intelligence[1]
Dept of Roman Contrastive Linguistics[2]
{vetulani,obrebski,jacekmar,gravet}@amu.edu.pl

META-NET

LTC 2013
Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November (???), 2013, Poznań, Poland

www.ltc.amu.edu.pl

contact: vetulani@amu.edu.pl



Adam Mickiewicz University in Poznań (UAM) is one of the most important in Poland. **(www.amu.edu.pl).**

The Faculty of Mathematics and Computer Science is one among the 15 faculties of the University.

The Department of Computer Linguistics and Artificial Intelligence exists since 1993.

Initial staff: Zygmunt Vetulani (head) and Krzysztof Jassem (assistant).

Staff and external collaborators
Zygmunt Vetulani (initiator and head of the department, professor)
Assoc. Professors: dr. J. Marciniak, dr. T. Obrębski
Collaborators: dr. J. Osiński, dr. G. Ligozat, prof. G. Vetulani
Ph.D. candidates/students: M. Kubis, G. Taberski

Ph.D. fellows: (graduated under the supervision of the Department): J. Daciuk, F. Graliński, M. Lison, J. Marciniak, T. Obrębski, J. Osiński, J. Walkowska

Long-term foreign visitors: G. Ligozat, G. Matlatipov, H. Madatov, M. Paprzycki, Y.Uetake

Emulated Language Competence and Other Selected Works in the Field of Polish Language Engineering

Natural language technologies are developped at the UAM in Poznań since many years. Some of these activities started in 70 and resulted with significant achievments (e.g. in the area of vocal synthesis (phonetisation) (prof. M. Steffen-Batóg)).

Activities in Computer Linguistics at the Faculty of Mathematics and Computer Science started after Z. Vetulani visited (1984) the GIA laboratory at the University Aix-Marseille II /Artificial Intelligence Group headed by Alain Colmerauer /(NL understanding, development of language resources) .

Emulated Language Competence and Other Selected Works in the Field of Polish Language Engineering

First implementations:

- The Polish Module to the ORBIS system (a bilingual /French and English/ Question-Answering interface created by Colmerauer and Kittredge in 1982). Probably the first non-trivial, natural language understanding interface for Polish. This system gave birth to the family of POLINT systems /still in progress/.
- Application: a NL- interface to Art History Data Base EXPEART (prototype) (within a CPBP 08.05; contract 3 05 01 09) (1991)
- Since 1993: development of successive versions of the POLINT family systems based on DCG and various optimization techniques: pre-analysis, switches, two-run processing, heuristic parsing. Effect: quasi-linear time complexity, low ambiguity processing. The practical effect is real time text processing in the systems with NL understanding interface (as e.g. POLINT-112-SMS /2006-2010/.

SOME PUBLICATIONS

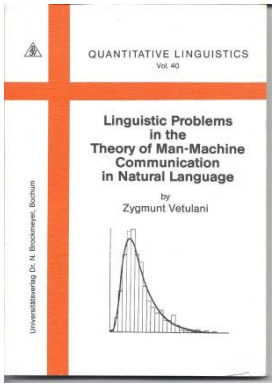
- PROLOG Implementation of an Access in Polish to a Data Base, in: Studia z automatyki, XII, PWN, 1988, p. 5-23.

- Z. Vetulani (2004): Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej, Akademicka Oficyna Wydawnicza EXIT, Warszawa.

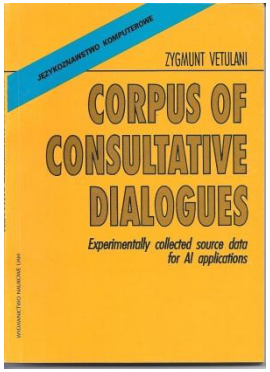
- An Expert system for Art History Data and Documents, in: Jerzy Bańczerowski (ed.), The Application of Microcomputers in Humanities, Adam Mickiewicz University Press, Poznań, 1991 (ISBN 83-232-0291-1), p. 63-74 (co-author J. Martinek).

Research on Dialogue Modeling financed by the Alexander von Humbold Foundation (in form of research grant awarded to Zygmunt Vetulani from 1987 to 1989):

- Linguistic problems in the theory of man-machine communication in natural language. A study of Consultative Question-Answering Dialogues. Empirical Approach. Brockmeyer, Bochum, 1989, (p. 150).

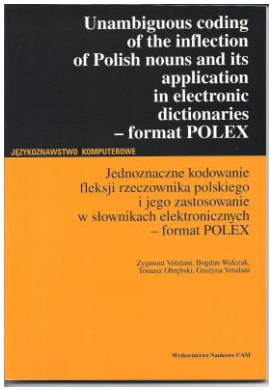


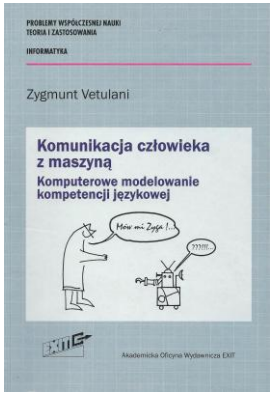
- Corpus of Consultative Dialogues. Experimentally collected source data for AI applications. Adam Mickiewicz University Press, Poznań, 1990 (p. 189).



Some early projects in LRs:
POLEX – Polish Lexical Data Base (1994–1996)

The main objective of POLEX was to create morphological electronic dictionaries for the core Polish vocabulary of general interest, based on a precise machine-interpretable formalism and operational coding system. We also considered as important factors human transparency and readability of the formalism. This aspect, sometimes considered as secondary from the automatic processing point of view, is however important for maintenance and further development of electronic resources (openness).





Dates: 1995–1996 (grant KBN8S50301007)

Over 100.000 entries

Distributed through ELRA/ELDA (free for research)

POLEX is the core of the UNIX based text processing environment UAM TEXT TOOLS

POLEX

POLEX morphological entries have the following shape:

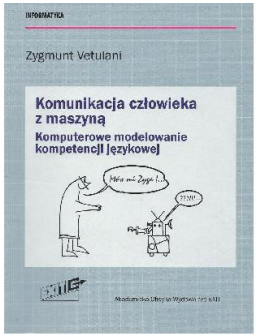
BASIC_FORM +
LIST_OF_STEMS +
PARADIGMATIC_CODE +
STEMS_DISTRIBUTION

For example, the dictionary items for *frajer*^{I} and *frajer*^{II} will be as follows:

frajer; frajer,frajerz; N110; 1:1-5,9-13; 2:6-8,14
frajer; frajer,frajerz; N110; 1:1-5,8-14; 2:6-7

Some early projects in LRs: CEGLEX (Copernicus 1032)

The main goal of the CEGLEX consortium was to test the GENELEX proposal of a generic model for re-usable lexicons /first implemented for a numbe of West-European languages (French, English, German, Itralian,...)/ for three more languages: Czech, Hungarian and Polish. This generic model takes the form of a SGML DTD where linguistic information is associated with words via SGML tags and attributes. **Main deliverable**: prototype of 3-layered /morphology, syntax, semantics/ lexicon-grammar for based on the GENELEX format Polish (implemented for 3000 entries).



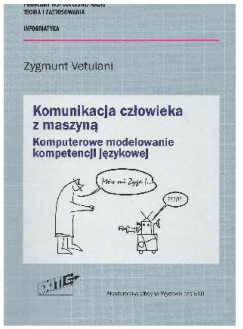
Dates: 1995-1996

Partners:

- GSI-ERLI, Charanton, France, Antoine Ogonowski (European Coordinator);
- Adam Mickiewicz University, Zygmunt Vetulani (Polish Coordinator);
- Lingware, Pecs, Hungary, Karoly Fabricz;
- Charles University, Prague, Czech Republic, Jan Hajicz.

Some early projects in LRs: GRAMLEX (Copernicus 621)

The aim of the COPERNICUS Project 621 GRAMLEX was to facilitate the initiation, coordination and standardization of the construction of morphological dictionary packages : French, Hungarian, Italian and Polish, including detailed formal description of the morphology of the languages. A morphological dictionary package is the set of lexicons and programs (as e.g. recognizer and a generator of inflected forms). **Main deliverables**: morphologic digital dictionaries (SGML coded).



Dates: 1995-1998

Partners:

- ASSTRIL, Marne - la - Vallée, France, Eric Laporte (European Coordinator);
- Adam Mickiewicz University, Zygmunt Vetulani (Polish Coordinator);
- Hungarian Academy of Sciences, Budapest, Hungary, headed by Julia Pajz;
- Morphologic, Budapest, headed by Gabor Proszeki;
- Consorzio Lexicon Ricerche, Salerno, Italy, headed by Mario Montoleone.

MAIN PUBLICATIONS:

- Vetulani, Z., Walczak, B., Obrębski, T., Vetulani, G.: Unambiguous coding of the inflection of Polish nouns and its application in the electronic dictionaries – format POLEX / Jednoznaczne kodowanie fleksji rzeczownika polskiego i jego zastosowanie w słownikach elektronicznych – format POLEX, Adam Mickiewicz University Press, Poznań, 1998
- Vetulani, Z., Martinek, J., Vetulani, G.: A description of Lexical Knowledge for Polish within the Genelex Model, in: K. Sroka (ed.) Kognitive Aspekte der Sprache. Akten des 30. Linguistischen Kolloquiums, Gdańsk, 1995, Max Niemeyer Verlag, Tübingen, 1996, pp. 175–180.
- Vetulani, Z., Martinek, J., Vetulani, G.: The CEGLEX dictionary model for Polish, in: R. Bazylewicz, O. Kossak (eds.), Proceedings of the 4th and 5th International Conferences UKRSOFT (Lviv, 1994, 1995), ISBN 5–7773–0338–2, SP «BaK», Lviv, 1995, pp. 144 – 150
- Vetulani, Z., Ligozat, G., Marciniak, J., Martinek, J.: Modelling of linguistic competence for guiding a robot: a corpus based approach, IJCAI–97, in: H. Guesgen (ed.) Spatial and Temporal Reasoning, Nagoya Congress Center, Japan, 1997, 19–23
- Vetulani, Z., Martinek, J., Obrębski, T., Vetulani, G. : Lexical Resources and Tools for Tagging Polish Texts within GRAMLEX, in: Investigationes Linguisticae, XXI:2, 1997, 401–416.
- Vetulani, Z.: Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX. In: M. Gavrilidou et al. (eds.), Second International Conference on Language Resources and Evaluation, Athens, Greece, 30.05.–2.06.2000, (Proc.), ELRA, pp. 367–37[illegible]

GRAMMARS AND PARSING

Contributions to the NL formal description tools for free word order languages: FROG grammars

- Z. Vetulani: Free Order DCG (FROG) in application to formal description of semi-free order languages, in: Kłopotek M., Tchórzewski J. (eds.), *Sztuczna Inteligencja, Materiały V Konferencji Naukowej*, Wyd. Akademii Podlaskiej, Siedlce, 2002, pp. 59-72.

- Z.Vetulani and F. Graliński (2005): Reinterpreting DCG for free word order languages, *Archives of Control Sciences*, vol. 15 (LI), No. 4 2005, 691-702.

ONTOLOGIES

Formal ontologies are considered as means of "formalisation of conceptualisation" (Gruber).

Ontologies may serve as reasoning support because of their mathematical structure, e.g. hierarchies of concepts which e.g. permit to implement the *default reasoning* and *inheritance.*

For our NL applications we have made the choice in favor of **WordNet-like** ontologies.

The term "WordNet" refers to the lexical base created at the Princeton University (1985, George A. Miller) also known as the Princeton WordNet (PWN).

ONTOLOGIES

- **Ontological problems related to construction of natural language interface for a mobile robot,** in: H. Guesgen (ed.): Workshop on Hot Topics in Spatial and Temporal Reasoning, IJCAI'99 (proceedings), 1999, Stockholm. pp. 31-36. (co-author: J. Marciniak).

- **Ontology of Spatial Concepts in Natural Language Interface for a Mobile Robot**, Applied Intelligence 17, Kluwer Academic Publishers, 2002, pp. 271-274. (co-author: Marciniak, J.)

- **Linguistically Motivated Ontological Systems**, in: Callaos, N, Lesso, W., Schewe K.-D., Atlam, E. (eds.): Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, July 27- 30, 2003, Orlando, Florida, USA, vol. XII (Information Systems, Technologies and Applications: II), Int. Inst. of Informatics and Systemics, 2003, pp. 395-400.

- Z. Vetulani (2004): **Towards a Linguistically Motivated Ontology of Motion: Situation Based Synsets of Motion Verbs.** In: Barr, V., Markov, Z. (eds.) Proceedings of the Seventheens International Florida Artificial Intelligence Research Society Conference (FLAIRS-04), AAAI Press (2004), Menlo Park, California, 813-817.

PolNet–Polish Wordnet

PolNet (the full name of the project is "PolNet–Polish Wordnet") started in 2006 and was first intended to serve as ontology for the POLINT–112–SMS system.

At the project start in 2006 there was no available wordnet for Polish. Now, PolNet is free distributed for non–commercial usage (version v1.0) under the Creative Commons license:

Creative Commons Attribution–NonCommercial–NoDerivs 3.0 Unported License.

PolNet–Polish Wordnet

Initially, PolNet (v.0....) was containing only the noun synsets composed of simple words.

Now

- **Nouns** : about **11,700 synsets** for **20,300 word-sens pairs** (for **12,000** common nouns).
- **Verbs** : in 2011, the a **verbal part** of PolNet consisted of env.**1,500** synsets, (for **900** verbs).

The verbal part is in development. Works in order to include compound nouns (in particular the verb-noun collocations) are also in an advanced phase.

(For reference we bring the reader's attention to the fact that the **basic vocabulary** sufficient to satisfy the needs of ordinary, every-day conversation has been evaluated for 1000-2000 mots. According to Ogden /1930/ the size of the Basic English is about 850 words)

WORDNETS

- A wordnet is an ontology composed of concepts represented as **classes of synonyms (synsets)** and related by conceptual relations induced from relations holding between words. The first one is (Miller, since 1985)

- A wordnet composed of **noun-derived and verb-derived concepts** may (and should) include relations **encoding the valency** of predicative words (i.e. the grammatical information associated to these words)

WORDNETS

The main idea is **simple**:

to gather together synonyms into equivallence classes (which represent concepts) and to consider relations holding between these classes.

In practice, this idea is **difficult** to implement because of the phenomenon of **word polysemy**. One has to consider *disambiguated words* instead of *words* (or more precisely *word+word_sense pairs*).

The equivallence classes of synonymous, disambiguated words are called **synsets.**

PolNet–Polish Wordnet

The concepts represented in a wordnet through synsets correspod to the way the conceptualisation is reflected in the language.

As it is well known that conceptualisation of the world may not be the same for speakers of different languages the propoer way to create a new wordnet is

to do it from scratch.

This was the solution chosen for PolNet (the full name of the project is "PolNet–Polish Wordnet") .

For quality reasons most of the work is to be done under human control (computer aided) .

PolNet–Polish Wordnet

The PolNet synsets are linked by relations. The two main relations are ***hyponymy*** et ***hyperonymy*** for noun synsets and ***semantic roles*** for verbal synsets.

The **selection of concepts** to be represented in PolNet was done on the basis of **word frequences** observed in the National Corpus of Polish (the IPI PAN, Przepiórkowski) and in small expermental (domain-oriented) corpora collected within the POLINT-112-SMS project.

The word meaning identification was done manually on bases of traditional dictionaries of Polish. Also **synsets were set manually** by lexicographers assisted by a specialised software, namely the DEBVisDic system made at the Masaryk University of Brno (Czech Republic /Pala, Rambousek/).

PolNet–Polish Wordnet

First step: for simple nouns

PolNet was developed **from scratch** on basis of linguistic knowledge encoded in **dictionaries** and therefore reflects the conceptualisation of the world proper to the Polish language speakers.

- With one exception (for testing purposes), PolNet is based on **the most frequent words** and therefore may be considered as representative for the language core.
- Initial PolNet:
  - some 11,700 synsets for over 20,000 word-meaning pairs (for 12,000 nouns)
  - organizing feature : hyponymy/hyperonymy

PolNet–Polish Wordnet

- EXAMPLE: synset {szkoła:1, buda:5, szkółka:1,...}
- (generated under DebVisDic)

- <SYNSET>
- <ID>PL_PK-518264818</ID>
- <POS>n</POS>
- <DEF>instytucja zajmująca się kształceniem; *educational institution* </DEF>
- <SYNONYM>
- <LITERAL lnote="U1" sense="1">**szkoła**</LITERAL> % *szkoła=school*
- <LITERAL lnote="U1" sense="5">**buda**</LITERAL>
- <LITERAL lnote="U1" sense="1">**szkółka**</LITERAL>
- .....
- </SYNONYM>
- <USAGE>Skończyć szkołę</USAGE>
- <USAGE>Kierownik szkoły</USAGE>
- .....
- <**ILR** type="**hypernym**" link="POL-2141701467">instytucja oświatowa:1</ILR>
- <**RILR** type="**hyponym**" link="POL-2141575802">uczelnia:1,szkoła wyższa:1,wszechnica:1</RILR>
- <**RILR** type="**hyponym**" link="POL-2141603029">szkoła średnia:1</RILR
- .....
- <STAMP>Weronika 2007-07-15 12:07:38</STAMP>
- <CREATED>Weronika 2007-07-15 12:07:38</CREATED>
- </SYNSET>

Second step: wordnet for simple verbs

- preparative action
  - selection of 1,530 verbs (selected as we did for nouns)
  - creation of a valency dictionary of simple verbs (VDSV) (input from: existing dictionaries incuding Polański dictionary of verbs)

- creation of verbal synsets in PolNet starting with selected verbs (900 of 1,533)
  - over 1500 resulting synsets for 2,900 word-meaning pairs (for 900 most important verbs)
  - organizing feature: **semantic role** relations

PolNet–Polish Wordnet

Synset: {pomóc:1, pomagać:1}

- XML presentation of the same synset:
- <SYNSET>
- <VALENCY>
- <FRAME>Agent(N)_Benef(D)</FRAME>
- <FRAME>Agent(N)_Benef(D) Action('w'+L)</FRAME>
- <FRAME>Agent(N)_Benef(D) Manner</FRAME>
- <FRAME>Agent(N)_Benef(D) Action('w'+L) Manner</FRAME>
- </VALENCY>
- <ILR type="category_domain" link="1356">CITTA:1</ILR>
- <ILR type="Agent" link="ENG20-02383992-n">człek:1, człowiek:1, istota ludzka:1, zwierzę:2, ...</ILR>
- <ILR type="Benef" link="ENG20-02383992-n">człek:1, człowiek:1, istota ludzka:1, zwierzę:2, ...</ILR>
- <ILR type="Action" link="PL_PK-2035015933">czynność:1</ILR>
- <ILR type="Manner" link="2214">CECHA_ADVERB_JAKOŚĆ:1</ILR>
- <DEF>"wziąć (brać) udział w pracy jakiejś osoby (zwykle razem z nią), aby ułatwić jej tę pracę"</DEF>
- <SYNONYM>
- <WORD>pomóc</WORD>
- <WORD>pomagać</WORD>
- <LITERAL lnote="U1" sense="1">pomóc</LITERAL>
- <LITERAL lnote="U1" sense="1">pomagać</LITERAL>
- </SYNONYM>
- <ID>3441</ID>
- <USAGE>Agent(N)_Benef(D); "Pomogłam jej."</USAGE>
- <USAGE>Agent(N)_Benef(D) Action('w'+L); "Pomogłam jej w robieniu lekcji."</USAGE>
- <USAGE>Agent(N)_Benef(D) Manner Action('w'+L); "Chętnie pomogłam jej w lekcjach."</USAGE>
- <USAGE>Agent(N)_Benef(D) Manner;"Chętnie jej pomagałam."</USAGE>
- <CREATED>ag[illegible]2010-11-27 18:49:47</CREATED>
- <POS>v</POS>
- </SYNSET>

Verb-Nouns Collocations

Long term verb-noun collocations projects

Systematic studies of verb-noun collocations were initiated in the late 1990s by Grażyna Vetulani (2000). The first step consisted in manual examination of env. 40,000 of Polish nouns resulting in extraction of about 8,000 abstract predicative nouns.

Then 5 different classes were identified. The first one, the most irregular, was analysed and described in a format which permits to encode its valency structure (including the required, semantically empty (or almost empty), *support verb.*

Project *"Rozbudowa zasobów cyfrowych języka polskiego w zakresie słowników walencyjnych w kierunku leksykonu-gramatyki zorientowana na potrzeby zastosowań informatycznych w humanistyce" ("Lexicon-grammar oriented extension of Polish digital valency dictionaries for computer applications in humanities"),* febr. 2012-febr. 2015, Grant nr: 11H11 010080 / Numer umowy: MNiSW Nr 0022/FNiTP/ H11/80/2011 (managed by Grażyna Vetulani; Faculty of Modern Languages)

Verb-Nouns Collocations

- Example
- collocations found in dictionaries
- ambicja, f/ *(ambition)*
- mieć(Acc)/N1(Gen), *"mieć ambicję" (to have an ~ of sth)*
- mieć(Acc,pl)/MOD, *"mieć ambicje" (to have MOD ~s/)*

- collocations retrieved in the corpus[1] – manual analysis of concordances [2]
- ***posiada*ć(Acc,pl)/MOD, *"posiadać ambicje" (to own MOD ~s)***
- ***ujawnia*ć(Acc,pl)/MOD, *"ujawniać ambicje" (to show MOD ~s)***
- ***zaspokoi*ć(Acc)/N1(Gen), *"zaspokajać ambicję" (to fulfill one's ~ of sth)***
- ***zaspokoi*ć(Acc,pl)/MOD, *"zaspokoić ambicje" (to fulfill MOD ~s)***
- ***zaspakaja*ć(Acc)/N1(Gen) *"zaspakajać ambicję"(to fulfill one's ~ of sth)***
- ***Zaspakaja*ć(Acc,pl) *"zaspakajać ambicje"(to fulfill MOD ~s)***

- [1]IPI PAN CORPUS (80 mln), [2]generated using UAM TEX TOOLS

Third step: wordnet extended to **verb–noun collocations (and other compouned predicative lexemes)**

In order to extend Word Net and to make it more appropriate for applications a task was defined within the collocation project.

This task consists in enlargement of the resource to another class (Class II – feature names) and integration of the collected collocations with the wordnet.

PolNet–Polish Wordnet –evaluation

The computer–aided manual processing supported by DEBVisDic permitted us to obtain the **quality impossible to reach in the wordnet systems done entierly or mainly automatically / statistically** . The quality is however to be payed at high cost of human experts work and verification.

The **PolNet evaluation** was done at 3 levels :

- on–line manual evaluation at the coding time,
- with the help of a software tool (WQuery, Kubis)
- and within the POLINT–112–SMS application.

Publications

- 2000: *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych na tle porównawczym,* Wydawnictwo Naukowe UAM, Poznań.
-
- G. Vetulani, Z. Vetulani (2005): Kilka postulatów dotyczących słownika kolokacji werbo-nominalnych, w: D. Stanulewicz, R. Kalisz, W. Kürschner and C. Klaus, *De lingua et litteris: Studia in honorem Casimiri Andreae Sroka*, Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego. (ISBN 83-7326-267-9).
- Vetulani, G., Vetulani Z. Obrębski, T. (2006) : Syntactic Lexicon of Polish Predicative Nouns, N. Calzolari (ed.), *Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24-26.05.2006,* (Proceedings), ELRA, Paris, 1734-1737.
- 2012: *Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I*, Wydawnictwo Naukowe UAM, Poznań.

TIME and SPACE

1995-1997: French-Polish Joint Research of the Polish Government (KBN) and the French Government : "Accès en langue naturelle aux bases de connaissances spatiales", (ref 294/96/97 and 76157), domain: artificial intelligence, partners: UAM and University Paris XI (The project resulted with long term collaboration with Univ. Paris XI and LIMSI /Ligozat/)

Continuation

- Gerard Ligozat, Zygmunt Vetulani (2009) Reasoning About Events: the Spatio-Temporal XRCD Calculus, in: Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.), After Half a Century of Slavonic Natural Language Processing. Masaryk University. Brno. 231-245.

- Ligozat, G., Vetulani, Z., Osiński, J. (2011): Spatiotemporal Aspects of the Monitoring of Complex Events for Public Security Purposes, [In:] Spatial Cognition and Computation: An Interdisciplinary Journal, vol. 11 (1), str. 103-128, Taylor & Francis Group.
  http://www.tandfonline.com/doi/abs/10.1080/13875868.2010.544050#preview

- Gerard Ligozat, Zygmunt Vetulani (2009) Reasoning About Events: the Spatio-Temporal XRCD Calculus, in: Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.), After Half a Century of Slavonic Natural Language Processing. Masaryk University. Brno. 231-245

LEXICON GRAMMAR PROGRAM

A long term research program with the objective of organizing the grammatical description of Polish in a lexicon-grammar.

LEXICON GRAMMAR

= **a formalised NLgrammar** where

- elementary sentence is the fundamental unit of meaning
- and where **possibly complete** grammatical information is stored *together* with words
  (this helps organizing the grammatical knowledge)

This last feature is essential for NLP: **lexicon-grammars are computer-friendly**

There are various applications of lexicon-grammars (in text analysis, correction, disambiguation,...) but

our primary motivation was its **utility to enhance parsing** through identification of the (main) predicative element of the sentence to restrict the *parsing search space*

(implemented in POLINT systems, since the 1980s)

LEXICON GRAMMAR PROGRAM

A long term research program with the objective of organizing the grammatical description of Polish in a lexicon-grammar.

LEXICON GRAMMAR

= **a formalised NLgrammar** where

- elementary sentence is the fundamental unit of meaning
- and where **possibly complete** grammatical information is stored *together* with words
  (this helps organizing the grammatical knowledge)

This last feature is essential for NLP:

lexicon-grammars are computer-friendly

various reasons to develop lexicon-grammars
(text analysis, correction, disambiguation,...)

Our steps towards a Lexicon Grammar

- initial steps within CEGLEX,
- through partial practical implementations within POLINT
- systematic syntactic description of verb-noun collocations
- integration of grammatical information with PolNet

LEXICON GRAMMAR PROGRAM

OUR STEPS TOWARDS THE LEXICON GRAMMAR FOR POLISH

- initial steps within CEGLEX,
- through partial practical implementations within POLINT
- systematic syntactic description of verb-noun collocations
- integration of grammatical information with PolNet

OUR LEXICON GRAMMAR USE CASE

Preprocessing in the POLINT systems:

Heuristics built at preprocessing stage in order to controll parsing :

- structural hypothesis about the sentence syntax (to make parsing (more) deterministic)

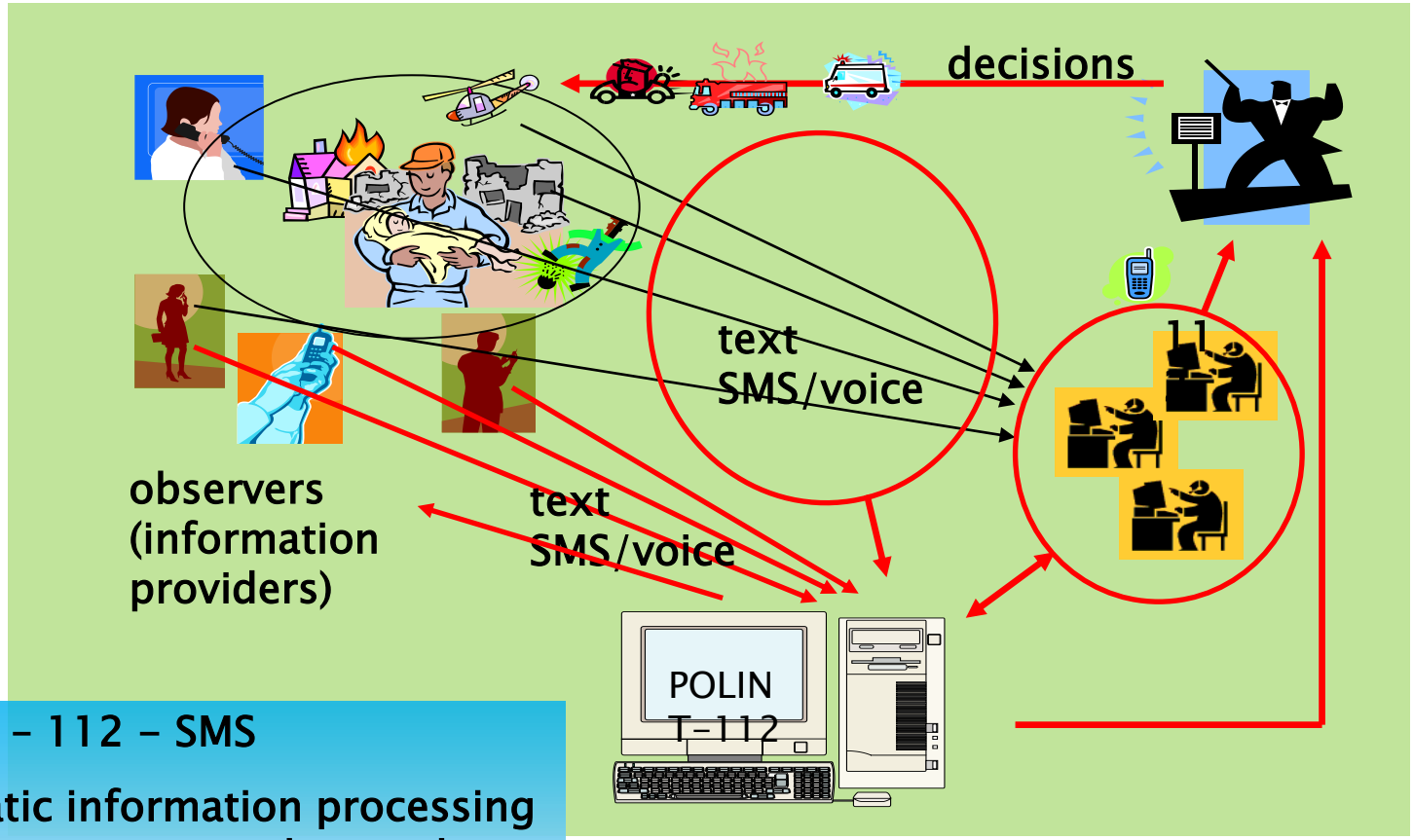
LEXICON GRAMMAR PROGRAM

Projects involved in the Lexicon-Grammar Development Program

1. Polex (morphology/inflection)
2. PolNet
3. POLINT systems (POLINT-112-SMS)
4. CITTA (City Tour Assistant)
5. Verb-Nouns Collocation projects

Supported by:
Polish Government (1,2,3,4,5), Citty of Poznań (3),
UAM (3, 4,5), Polish Platform for Homeland Security (3)

Application context : POLINT-112-SMS project



POLINT – 112 – SMS

automatic information processing involving NL-text understanding, information integration, contradiction solving, decision assistance

LEXICON GRAMMAR PROGRAM

Some publications

- Vetulani, G., Vetulani Z. Obrębski, T. (2006) : Syntactic Lexicon of Polish Predicative Nouns, N. Calzolari (ed.), Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24-26.05.2006, ELRA, pp. 1734-1737.

- Vetulani, Z., Obrębski, T., Vetulani, G. (2007): Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora. In Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference (FLAIRS-07), AAAI Press (2007), Menlo Park, California, 267-268.

- Vetulani, Z. (2012): Wordnet Based Lexicon Grammar for Polish, in: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, ‚Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 23-25, 2012, Istanbul, Turkey, ELRA, Paris (accessible through http://www.lrec-conf.org/proceedings/lrec2012/index.html)

- Vetulani, G. and Vetulani Z. (2012): Dlaczego Leksykon-Gramatyka?, w: Anna Dutka-Mańkowska, Anna Kieliszczyk, Ewa Pilecka (red.), Grammaticis unitis. Mélanges offers à Bohdan Krzysztof Bogacki, Wydawnictwa UW, pp. 308-316.

- Vetulani, Z., Vetulani, G. (in print): Through Wordnet to Lexicon Grammar. W: Fryni Doa et al. (eds.) Proceedings of the 30th Conference on Lexis and Grammar, 5-8.10.2011, Nicosia, Cyprus, Éditions Honoré Champion (Full text)

Parsing in POLINT-112-SMS

Technologies which contribute to understanding are **parsing** as well as **discourse analysis.**

In the POLINT-112-SMS system **parsing** is executed by the PROLOG interpreter of a properly formalized grammar (e.g. CFG /context free grammar/ DCG).

(PROLOG is a high level programming language based on logic, its interpreter may be considered as a shell of expert systems).

The main drawback: PROLOG may be ineffective. Our solution : heuristic parsing, where the main module (expensive when backtracking) is preceded by the pre-analysis (cheap) which simplify the input and generate heuristics whose role is to control the parsing execution (reduction of indeterminism).

(By "heuristics" we mean a procedure which guides the parser to make correct choices)

Parsing in POLINT-112-SMS

Correct and effective parsing requires application of several lowel level technologies to perform :

segmentation (sentences and words),
lemmatisation,
spell checking,
simplification,
desambiguation,
named-entity recognition.

In many cases, complete understanding is not possible on the basis of syntactic analysis alone (some context is to be taken into consideration).

Parsing in POLINT-112-SMS

Parsing, as well as generation, depends on basic resources which are **grammars** and **dictionaries**.

The grammars POLINT were integrated end directly applied in the project.

These grammars were elaborated for successive versions of question-answer systems POLINT produced since the 1990ties.

They are formally equivalent to the definite clause grammars (DCGs) directly translated into PROLOG.

They were then adapted in the way allowing them to be controlled by **heuristics** in order to minimize the non-determinisme of parsing. The result is that heuristics make parsing **executable practically in the linear time.**

POLINT dictionaries are of the kind of lexicon-grammars.

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

- On the basis of our *know-how* and technologies obtained so far we started in 2006 **a large project (POLINT-112-SMS) integrating several NL technologies.**
- This project was funded by Polish Goverment and was realised from 2006 to 2010 within a larger program (managed by Z. Vetulani), namely "Technologie przetwarzania tekstu polskiego zorientowane na potrzeby bezpieczeństwa publicznego" (Grant MNiSzW R0002802)/ PPBW
- Some of the tasks of this project are now continued within the project coordinated by Grażyna Vetulani (Faculty of Modern Languages) ("Lexicon-grammar oriented extension of Polish digital valency dictionaries for computer applications in humanities") (Grant MNiSW Nr 0022/FNiTP/H11/80/2011)

PPBW

Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego

Zygmunt Vetulani, Jacek Marciniak, Tomasz Obrębski, Grażyna Vetulani, Adam Dąbrowski, Marek Kubis, Jędrzej Osiński, Justyna Walkowska, Piotr Kubacki, Krzysztof Witalewski

WYDAWNICTWO NAUKOWE UAM

ISBN 978-83-232-2155-5
ISSN 1896-379X

Project team in 2010

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

The **main technology** of the project is **language understanding**. The system we have built (prototype) has as its main functionality to provide assistance in the decision making in critical situations which require an immediate action. Decisions must be taken on the basis of the current situation analysis. A representation of the current situationis is on-line compiled by the system from information elements and processed to obtain the decision supporting elements (finally arriving to the *Crisis Management Center (CMC)*.

As a study case we selected monitoring of the soccer stadium at a match observed by a large number of supporters. Such situation are usually generating a number of risks. Typically, such situations are covered by video-monitoring (often at100% coverage).

The police consultants of the project consider that the monitoring video (cameras) is unsufficient and the human, on-site supervising is necessary.

Problem: how to assure communication and how to complete interpretation of messages send by te informers.

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

Kowal nosi bluzę i czarne spodnie. Kowal i Wolski są na murawie. **Kowal i Wolski należą do bojówki.**

System: 'Zrozumiałem'

Osoba, która nosi czarne spodnie, jest niebezpieczna. Kowal i Wolski atakują policjantów kamieniami.

System: 'Zrozumiałem'

Kto atakuje policjantów?

System: 'Kowal. Wolski. Brak dalszych odpowiedzi!'

Kto jest niebezpieczny?

System: 'Kowal. Wolski. Brak dalszych odpowiedzi!'

Informator: czy istnieje niebezpieczna osoba, która nie została złapana?

System: 'Tak, kilka, w szczególności istnieje: Kowal !'

Kowal ucieka.

System: 'Zrozumiałem'

Kowal został złapany.

System: 'Zrozumiałem'

Co nosi osoba, która jest niebezpieczna, i która nie została złapana?

System: 'kamien. Brak dalszych odpowiedzi!'

W co jest ubrana osoba, która jest niebezpieczna?

System: 'spodnie czarne.bluzę. Brak dalszych odpowiedzi!'

Osoba, która ma czarne spodnie, wychodzi z stadionu.

System: 'Zrozumiałem'

Jak nazywa się osoba, która ma czarne spodnie, i która jest niebezpieczna?

System: 'Kowal. Brak dalszych odpowiedzi!'

Osoba, która nosi czarne spodnie, i która jest niebezpieczna,, wybiegła z stadionu.

System: 'Zrozumiałem'

Kto jest na murawie?

System: 'Wolski. Brak dalszych odpowiedzi!'

Wolski został złapany.

System: 'Zrozumiałem'

Kto nie został złapany?

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

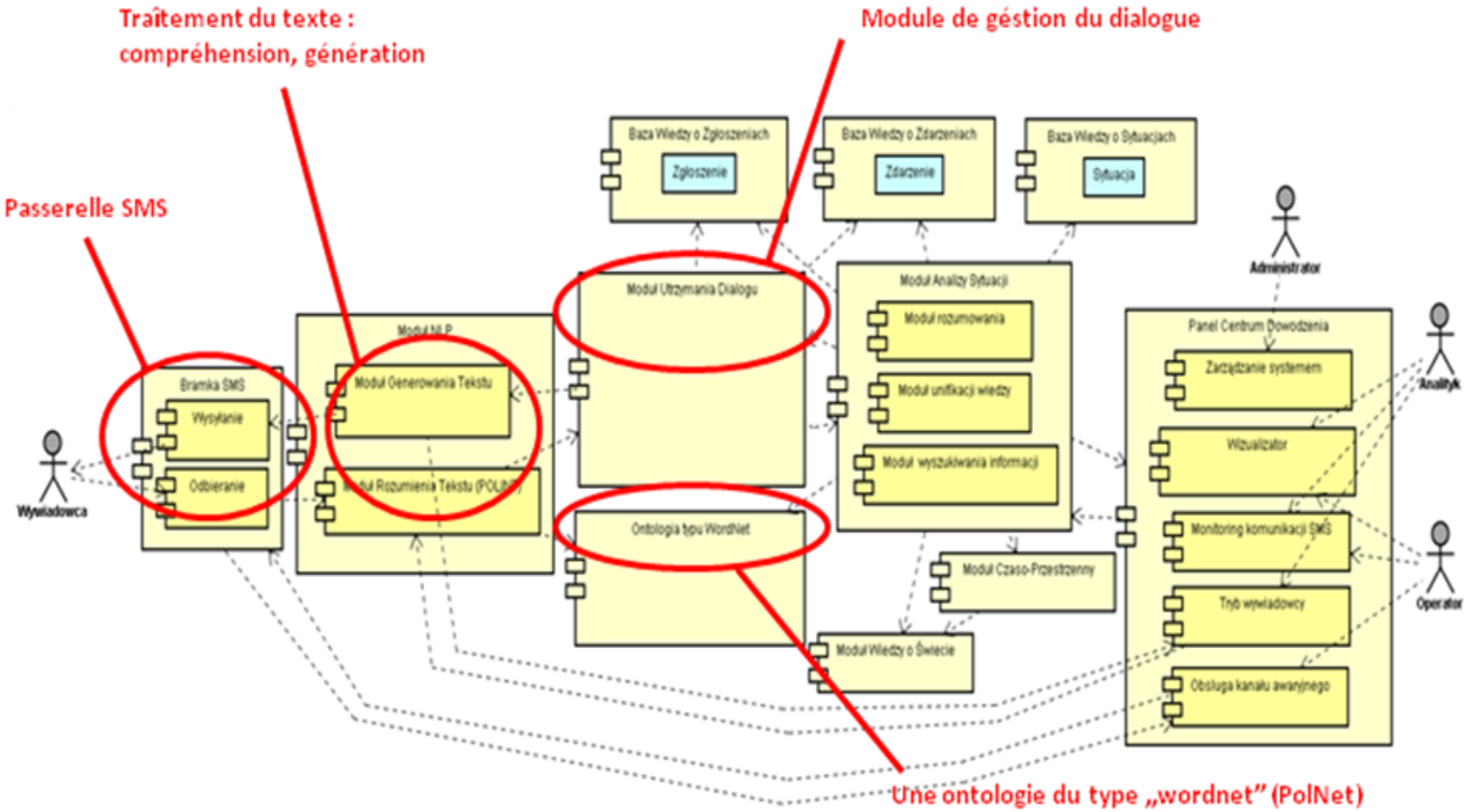
Message exchange is done in **natural (human) langage (Polish).** This means that the system must have **language competence** as well as **communicative competence.**

The prototype has been tested by public security experts both in simulated and real-life situations of a football match at the city stadium in Poznań.

Test messages (SMS) were exchanged using public cellular phones.

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

POLINT-112-SMS system architecture



In red: modules using natural language technologies.

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

a) SMS gate (capturing texte)

b) Natural Language Processing module

- understanding
- generation

c) Dialogue Maintenance Module

d) Situation Analysis Module

- desambuiguisation
- reasoning
- information search/query answering

e) Temporal analysis module d'analyse temporelle

f) Knowledge processing module

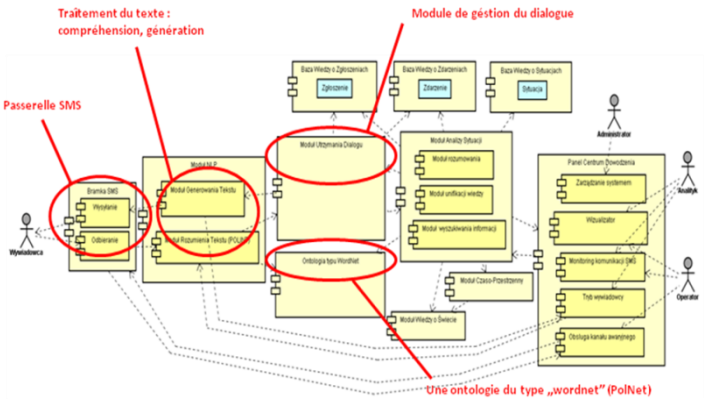
g) Ontology (PolNet)

h) Knowledge Bases

- about events
- about actes of commnication

i) CMC terminal (admin)

- visualisation
- administration
- capturing and displaying text



UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

POLINT-112-SMS system is a product of a **man-machine communication technology** which is an AI technology using in particular various (lovel level) natural language technologies.

The man-machine communication requires implementation of appropriate man-machine **interfaces**. In the case of POLINT-112-SMS we use NL text interfaces dedicated to two kind of users: information suppliers (informers) and target beneficients (CMC staff).

The informers' messages, queries and answers are entered to the system from mobile phones through the **SMS gate.**

The another input-output device is the terminal at the CMC. It recieves and outputs texts and displays the images recieved from the **visualisation submodule.** It is also possible to display the past dialogue in form of structured **text**.

UNDERSTANDING TECHNOLOGY: POLINT-112-SMS

POLINT-112-SMS system is a product of a **man-machine communication technology** (which is an AI technology).

Among the lovel level language technologies involved, the highest one is **understanding**. The NLP and Dialogue Maintenance moduls are both contributing to understanding.

The **understanding** software takes an element of the text and interprets it, i.e. it calculates its **representation** which is then submitted to further processing.

Typically, *the procedure of understanding a question* produces a formal object which initiates procedures responsible for *answer finding.*

Applications of wordnet based ontologies

- Applications in web portals/vortals and e-learning content repositories:
  - Support of content owners / editors whilst tagging resources
  - Search engines „semantically sensitive”

- Implementations in humanities:
  - Wilanów Palace Museum Vortal (Old Polish History and Culture wordnet based ontology; Marciniak 2011)
  - E-archaeology content repository (Archaeological heritage in contemporary Europe wordnet based ontology; Marciniak 2012)

E-learning content repository systems

Didactic content repository – database of e-learning content structured as learning objects

Content repository tool:

- Web software for sharing e-learning content
- Designed to attribute unambiguous didactic interpretations to didactically valuable parts of e-learning content
- Universal Curricular Taxonomy System (UCTS) used for interpretation of content (curriculum, module, unit) (Marciniak,2012)

E-archaeology content repository:

- E-learning content from the field of management and protection of archaeological heritage (ca. 5000 learning objects)
- *Archaeological heritage in contemporary Europe* wordnet based ontology used for tagging learning objects and UCTS elements

Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) – systems used in distance learning adapting to the needs of the learner

Research on content delivery strategies in huge and dynamic content repositories composed of learning objects using wordnet based ontologies

UAM TEXT TOOLS

UAM Text Tools (UTT) is a package of language processing tools developed (under UNIX) at the Department of Computer Linguistics and Artificial Intelligence on the basis of POLEX.

The main functionalities :

- tokenization
- dictionary-based morphological analysis
- heuristic morphological analysis of unknown words
- spelling correction
- pattern search
- sentence splitting
- generation of concordance tables
- syntactic parsing(umbiguous otput)
- semi-automatic morphological desambiguation

The toolkit is destined for processing of raw (not annotated) unrestricted text for any conceivable purpose.

The system is organized as a collection of simple command-line programs, each performing one operation, e.g. tokenization, lemmatization, spelling correction. The components are independent one from another, the unifying element being the uniform i/o file format. The components may be combined in various ways to provide various text processing services. Also new components supplied by the used may be easily incorporated into the system provided that they respect the i/o file format conventions.

UAM TEXT TOOLS

UAM Text Tools – some applications

- Principal projects, where UTT package was used:
- Verb-noun collocations dictionary (task: collocation extraction from corpora)
- Positive and negative symptoms of language disorder in schizophrenia (task: preparation of unambiguously tagged utterance transcriptions, computation of speech style parameters)
- POLINT-112-SMS (tasks: various tasks related to dialogue corpus analysis, SMS input spelling correction)
- Polish dependency grammar development enviromnent (morphological analysis)

UNIX-BASED TOOLS FOR LEXICON AND GRAMMAR DEVELOPMENT

pmdbsh (**P**olish **m**orphological **d**atabase **sh**ell) – an environment for Polex/PMDB maintenance and development

Polex/PMDB interface (browsing, modification)
export/update of UTT dictionaries
tools for semi-automatic Polex/PMDB extension
Ruby API to Polex/PMDB

dgsh (**d**ependency **g**rammar **sh**ell) – an environment for development of dependency grammar based on a large collection of examples (elementary phrases, complex phrases, sentences)
example database management
grammar development support tools

SOME OTHER UNIX-BASED TOOLS

Dependency parsers

dgp (**d**ependency **g**raph **p**arser) – all-paths dependency parser producing umbiguous output packed in the form of a dependency graph

ddp (**d**eterministic **d**ependency **p**arser) – single-path deterministic dependency parser with ability to revise (within certain limits) previously made head attachment choices according to a priority system based e.g. on dependency type ranking

META-NET

LTC 2013
Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November (???), 2013, Poznań, Poland

www.ltc.amu.edu.pl

contact: vetulani@amu.edu.pl

THANKS !