Grant Agency of the Czech Republic — Part C

Project description and substantiation

Applicant: Barbora Vidová Hladká

Title: An automatic coreference resolution procedure based on data collected by an alternative method (*PlayCoref*)

Project goals

This project aims to comprehensively solve the task of automatic coreference resolution in texts. The solution is comprised of training data preparation, the selection of a suitable machine-learning method, the implementation of the method and an evaluation of the results. An innovative method will be used for the preparation of training data, namely an Internet game whose design and implementation are also a part of this project. Czech and English will be used as the object languages.

Present state of the project

Applications of the "**human-machine communication**" type must be designed in such a way that working with them is as comfortable as possible for users. If possible, the communication should take place in a natural language for the user, or, even better, in the user's native language. Therefore, the ability of a machine to communicate in natural language must be based on a very detailed **knowledge of the language**.

Annotated corpora undoubtedly belong among the most important sources of "information" on natural language structure for its machine processing. Tools for practical language processing are built on this basis — in this way the corpora help bring the already attained knowledge of the language to the application area. Annotated corpora and tools based on them are parts of computational linguistics.

The preparation of annotated corpora is a demanding task in many senses — it is a time consuming and expensive activity. Let us exploit the capacity of Internet users who love to play games by offering them **on-line language games** with words, phrases or even whole documents — language games that will be fun for the players and at the same time provide the valuable data needed for machine-learning methods.

The establishment of the **LGame[1] portal**, offering on-line language games (Hladká, Ribarov, 2008), was motivated by the success of on-line games with images (van Ahn, Dabbish, 2004), which generate

[1] http://www.lgame.cz

data for the task of image processing — the ESP Game[2] (van Ahn, 2006) and Peekaboom[3] (van Ahn, Ruoran, Blum, 2006). This game portal offers progressive language games: By playing them, annotated corpora (data) are created for the tasks of natural language processing that have not yet been implemented (due to the lack of annotated data), or that have already been implemented but with very low performance. The first game that came into operation in the game portal was **Shannon's game.**[4] The experience from its design and implementation is invaluable for the design of all other games. All games presented in the game portal must satisfy 10 basic features:

1.	Playing the game generates data that is valuable for natural language processing tasks.	2.	During the game, the players are doing "something" until they achieve agreement.[5]
3.	Playing the game requires only a basic knowledge of the game language's grammar.	4.	The game rules are formulated independently of the game's language.
5.	The game is playable in Czech and English, at least.	6.	The game is of an interactive character: During the game, the players must have a general idea about what their opponent(s) do.
7.	The game is of a dynamic and quick character.	8.	The game is for at least two players.
9.	The game offers several levels of difficulty.	10.	Each game has a separate top score list.

The term **coreference**, as we understand it for the purposes of this project and as it was understood during the annotation of the Prague Dependency Treebank[6] (PDT 2.0) (Mikulová et al., 2006), is based on the term **reference**, i.e. on a relationship between terms and real world objects or situations. If two or more terms appear in a text referring to the same entity (person, object, phenomenon, fact), i.e. their reference is identical, their mutual relation is called a coreference. A sequence of respective terms in the text is called **a coreference chain**. In contrast with the annotation of the PDT 2.0, in our project we will work solely with endophoric references, which involve references within a text, i.e. we will not work with exophora, which involve references outside the text. Studies of coreference are a substantial part of discourse analysis. The following example is taken from the PDT 2.0; four coreference chains are marked in the text:

Bělorusko$_1$*: zastavení likvidace* ***arzenálů***$_2$

Moskva – Běloruský ***prezident***$_3$ *Alexandr* ***Lukašenko***$_3$ *nařídil pozastavit likvidaci vojenské* ***techniky***$_2$ *na území* ***republiky***$_1$*. Oznámil to v Minsku na čtvrtečním slavnostním večeru k oslavám Dne obránců vlasti. Opatření se týká tanků, letadel, obrněných transportérů a bojových vozidel pěchoty. Podle* ***Lukašenka***$_3$ *prý byl tento krok vyvolán ani ne tak nedostatkem finančních prostředků, jako spíše „patrným porušováním vytvořené rovnováhy sil ve světě“. Agentura Interfax soudí, že* ***prezident***$_3$ *měl na mysli přání východoevropských zemí vstoupit do Severoatlantické* ***aliance***$_4$*, což by pro* ***Bělorusko***$_1$ *znamenalo bezprostřední sousedství s* ***NATO***$_4$*.*

[2] http://www.espgame.org
[3] http://www.peakaboom.org
[4] http://www.lgame.cz/shannon
[5] Like in the ESP game, the players type in possible image labels until they agree on the same label.
[6] http://ufal.mff.cuni.cz/pdt20

Application that use natural language processing that are not limited to individual sentences and instead work with whole documents (e.g. machine translation, information retrieval, dialogue systems, systems for question answering) cannot manage without coreferences.

So far, worldwide attempts to solve the problem of coreference resolution in English texts work for simplicity with coreferences between name phrases. MUC-6[7] (Vilain et al., 1995; MUC-6, 1995) and MUC-7[8] (Hirschman, Chinchor, 1997) are the most frequently used methods for this type of data processing. The users will add to the text only the most basic information for coreference resolution. Many experiments use a discriminative approach — a supervised method — (Soon et al., 2001), where the task of coreference resolution is defined as a pair-wise discriminative task. The stratification of the task as a whole brings an increase of the number of algorithmic approach choices, whose list as well as their comparison is presented in Iida et. al, (2005). Uryupina, (2006) studies the impact of the choice of discriminative features on the quality of the results. Recently, non-discriminative approaches were also successful, e.g. Denis and Baldridge, (2007) reformulated coreference resolution as a task of integer linear programming, and Haghighi and Klein (2007) applied a combination of unsupervised machine learning with a Bayesian approach to the same task.

For the Czech language, coreference resolution was studied on tectogrammatical trees (representation of language meaning) from the PDT 2.0 (Kučová and Žabokrtský, 2005;, Nguy Giang, 2006; Němčík, 2006).

Formulation of the project's content and goals

1. Defining the goals of the project and the approaches to their solution

The project aims to comprehensively solve the task of **automatic coreference resolution in texts**. The solution is comprised of the preparation of training data, the selection of a suitable machine-learning method, the implementation of the method and an evaluation of the results. An innovative method will be used for the preparation of the training data, namely an Internet game CG game, whose design and implementation are a part of the project as well. Czech and English will be used as the object languages.

In the **CGame**, texts will be set before the players and they will mark words (expressions) that refer to the same entity (person, object, phenomenon, etc.). The design of the game will not require knowledge of linguistic terms from the players regarding coreferences. This game will become a part of the LGame portal: It will bring entertainment with a language component to the users.

The data acquired during the game will be used as training and test data for an automatic coreference resolution procedure. This procedure will apply the chosen method for coreference resolution and get

[7] http://cs.nyu.edu/faculty/grishman/muc6.html
[8] http://www-nlpir.nist.gov/related_projects/muc/

quantifiable results with standard evaluation measures — precision, recall and F-measure. As a part of the research, unsupervised methods (Haghighi and Klein, 2007), or optimisation methods of linear programming (Denis and Baldridge, 2007) will be compared with the various settings of supervised methods (Iida et. al, 2005), with an emphasis on the optimisation of parameter setting and feature selection, along with a subsequent error analysis.

The project covers data and tool components. We will specify the goals and the solution strategies for each of them separately:

DATA

- **The CGame**

Setting the rules of the game. The game rules will be formulated to satisfy the 10 basic features listed above. We stress that the players will not be loaded with any linguistic terms. Setting the rules covers the basic algorithm of the game, the scoring processes, the version for more than two players, and the graphic interface.

The selection of suitable Czech and English texts. Czech texts will be selected from the Prague Dependency Treebank 2.0 (PDT). This preference is motivated by the fact that the PDT 2.0 already consists of coreference annotations, restricted to the cases where the reference is rendered as pronoun; the annotation is built upon tectogrammatical trees (Mikulová et al., 2006). Such preference will bring very interesting comparisons of the coreference annotations directly upon texts (from the game) with the annotations upon the tectogrammatical trees. Similarly, English texts will be selected from the corpus with a manual coreference annotation that is based on the annotation scheme proposed by the MUC-6 and MUC-7 initiatives as a part of the coreference analysis. The choices for the English corpus will be researched in our project. The estimate of necessary text volume depends on the demands of the selected machine-learning methods.

Estimates of the game's popularity and its contribution. It is very hard to estimate how popular a text-based game will become. At least there exists a parallel with recently launched activity in the field of image processing, which has received a great deal of publicity from an Internet audience and has shown surprising results.[9] The situation with texts seems to be different in a sense. While observing images, one can easily list possible labels. In the case of a text, one must read it to detail its topics. Certainly, reading texts takes more time than observing images — the longer text, the more time consuming.

[9] *"A total of 13,630 people played the [ESP Game] during this time [August 9–December 10, 2003], generating 1,271,451 labels for 293,760 different images. Over 80% of the people played on more than one occasion. Furthermore, 33 people played more than 1,000 games (this is over 50 hours playing the game).*" (von Ahn, Dabbich, 2004). Over 10 million labels have been collected as of March 2008.

To deal with this issue we will produce goal-directed advertising and promotion for the game. When the first version of the CGame is released, students will play a certain number of sessions. We have a preliminary agreement with Seznam.cz concerning media support.

- **Data collected during the game sessions.** During the game's registration process, additional information regarding the user's year of birth, gender, and mother tongue will be collected mainly as a statistical survey. For each session, its results will be stored. The additional information and results will be stored in a selected database (e.g. PostgreSQL or MySQL). The data will then be exported from the database to the CSTS format, which is based on the SGML format and is widely used in computational linguistics to represent morphologically and syntactically annotated corpora.[10]

TOOLS

- **Implementation of the CGame .** The points defined in the specifications of the game will lead to their implementation: the choice of the programming language (bearing in mind PHP5 and Flash); functionality requirements (users' administration, controlling of the game's process, input data format, the export of collected data); system architecture; extensiveness; and security.
- **Graphical data browser ViewCoref.** A graphical browser ViewCoref will be created for a more comfortable analysis of the data collected during the game sessions. Coreference chains will be presented in an objective and pictorial way (by colors, indices, arrows, etc.). The browser will be implemented in Java.
- **Automatic coreference resolution procedure:**

Machine-learning method selection. The unsupervised method accompanied by the Bayesian sampling (Haghighi and Klein, 2007) will be set as the very first step of studying, implementing, and applying the data for improving the results. Moreover, we will select at least one state-of-the-art candidate from both the supervised and unsupervised methods. The implementation of all selected algorithms will possibly be enriched by their modifications for the sake of optimisation and the achievement of better results.

Specification of language-based information. No matter whether the supervised or unsupervised approach is used, each method requires good feature selection and extraction. We are supposed to use not only the basic statistical and frequency features of text units, but morphological features as well. We plan to use external tools for the morphological analysis and tagging (the disambiguation of the morphological analysis) for both Czech and English.

Implementation environment selection. The implementation of the selected methods eventually depends on the availability of high-quality and on the proven software or partial software libraries. During the steps such as text processing or training/test data preparation, we use the power and effectiveness of Perl. We will make every effort to use reputable computing libraries (e.g. to solve the two-class classification problem with classification trees or to sample from a particular distribution

[10] http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/csts/html/DTD-HOME.html

with help of the Bayesian approach) most likely accessible as packages in the statistical computing environment R or as C-libraries. We will write our own code (in R, or C eventually) for the still non-existent implementation of some subtasks. We put the focus on the modularity within the system and on the elaboration of user-friendly documentation. The final “package” for the coreference resolution task (i.e. the entire modular-based system composed of all atomic parts) will be freely available, including the collected data and examples.

Evaluation. Keeping in mind machine-learning principles, the collected data will be split into training and test data before entering it into a particular method. Given that, we will preserve a balanced and objective point of view. A thorough analysis of the results of all implemented methods and their comparison (using *precision, recall, and F-measure*) will be undertaken. We will conclude with the comparison of our results to the results of other relevant works.

2. Time Schedule

The project is proposed for 3 years. We provide a schedule for the data and tool components separately so that the subtasks listing corresponds to their mutual cohesion not only within one component but across the components:

DATA	TOOLS
First year	
Setting the rules for the Cgame.	
Czech and English texts preparation.	Specification of CGame implementation.
	Implementation of the CGame.
	ViewCoref browser implementation.
	Specification of the unsupervised method for an automatic coreference resolution procedure.
Second year	
Publication of the first version of the Cgame.	Unsupervised method implementation.
Test game sessions.	Cgame improvement based on the test game sessions.
Session data analysis.	Specification of the supervised method for an automatic coreference resolution procedure.
Third year	
Publication of the final version of the Cgame .	Supervised method implementation.
Publication of the data.	Publication of the automatic coreference resolution procedure with the highest performance.

3. Project Team

Barbora Vidová Hladká — project coordination; responsibility for expenditures; supervision of the data component.

Jiří Mírovský — supervision of the tool component; Cgame implementation; ViewCoref browser implementation.

Pavel Schlesinger — responsibility for the research, implementation and evaluation of the automatic coreference resolution procedure.

4. Results and their presentation

- The CGame will be presented on the LGame portal.
- The data collected through the Cgame will be at the disposal of researchers both in the Czech Republic and abroad.
- An automatic coreference resolution procedure will be freely available.
- The results will be presented in the form of contributions at international and national conferences relevant to the topics.
- The ongoing results will be published on the project’s homepage.
- Meetings will be organised with students to popularise the linguistically oriented game portal.

5. Application of the results

The project systematically aims at the coreference resolution task within natural language processing. The coreference resolution must be a part of the applications that work beyond the sentence border, i.e. that work with documents. Within Czech-language processing, the systematic solution of this task has not yet been provided. At the same time, the proposed alternative manner of collecting annotated data has not yet been applied to any language. The independence of the game rules on the language under consideration opens the possibility for the inclusion of other languages aside from Czech and English. The data collected during the game sessions will be at the disposal of other research projects. We will arrange for the MUC community to provide this data for international competition in question-answering systems (information extraction, coreference resolution, and named entity identification).

Technical support for the project

This project will be carried out at the Institute of Formal and Applied Linguistics (MFF UK), which has at its disposal the suitable hardware and software.

References

von Ahn, Luis. 2006. Games with A Purpose. *IEEE Computer Magazine* 39 (6), pp. 92-94.

von Ahn, Luis, Ruoran Liu, Manuel Blum. 2006. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the ACM CHI 2006*.

von Ahn, Luis, Laura Dabbish. 2004. Labeling Images with a Computer Game.In *Proceedings of the ACM CHI 2004*, pp. 319-326.

Denis P., Baldridge. 2007. Global, joint determination of anaphoricity and coreference resolution using integer programming. *Proceedings of HLT-NAACL*, Rochester: pp.236–243.

Haghighi A, D. Klein. 2007. Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague:pp. 848–855.

Hirschman Lynette, N. A. Chinchor. 1997. MUC-7 Coreference Task Definition. *Proceedings of MUC-7*.

Hladká Barbora, Kirila Ribarov. 2008. Play the Language: An Alternative Manner of Collecting Annotated Data, submitted for the ACL 2008, Columbus, Ohio, USA.

Iida R., K. Inui, Y. Matsumoto. 2005. Anaphora resolution by antecedent identification followed by anaphoricity determination. Transactions on Asian Language Information Processing (TALIP), 4(4):417–434.

Kučová Lucie, Zdeněk Žabokrtský. 2005. Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution. In Matoušek, V., Mautner, P., Pavelka, T. (Eds.): *Text, Speech and Dialogue*, 8th International Conference, TSD 2005, Karlovy Vary.

Mikulová Marie, Bémová Alevtina, Hajič Jan, Hajičová Eva, Havelka Jiří, Kolářová Veronika, Kučová Lucie, Lopatková Markéta, Pajas Petr, Panevová Jarmila, Ševčíková Magda, Sgall Petr, Štěpánek Jan, Urešová Zdeňka, Veselá Kateřina, Žabokrtský Zdeněk. Anotace na tektogramatické rovině Pražského závislostního korpusu. Referenční příručka.*Tech. Report 31 ÚFAL MFF UK*, 2006, 183.

MUC-6. 1995. Proceedings of the Sixth Message Understanding Conference (MUC-6), November 1995, SanMateo: Morgan Kaufmann.

Němčík Václav. Anaphora Resolution.2006. Master thesis MUNI, Brno.

Nguy Giang Linh. 2006. *Návrh souboru pravidel pro analýzu anafor v českém jazyce*. Diplomová práce MFF UK, Praha.

Soon W., H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Uryupina O. 2006. Coreference Resolution with and without Linguistic Knowledge. *LREC Proceedings*:993–898.

Vilain Marc, et al.1995. A Model-Theoretic Coreference Scoring Scheme, Proceedings of the SixthMessage Understanding Conference (MUC-6), pp. 45-52, November 1995.

Czech Science Foundation – Part D

Applicant

Applicant: Barbora Vidová Hladká

Curriculum vitae: Barbora Vidová Hladká

Date of Birth: September 2, Hlinsko hladka@ufal.mff.cuni.cz
Address: Semilská 926, Prague 9 – Kbely, 197 00

Affiliation:
Faculty of Mathematics and Physics (MFF), Charles University in Prague (UK), Ke Karlovu 3, 121 16 Prague 2
Department: Institute of Formal and Applied Linguistics MFF UK, Malostranské nám. 25, 118 00 Prague 1

Education:
1989–1994 MFF UK, computer science (MSc degree)
1994–2001 MFF UK, computational linguistics (PhD degree)

Professional experience:
1996–present: Institute of Formal and Applied Linguistics (Charles University in Prague) – corpus and computational linguistics, machine learning

Important publications from the last five years – books

Barbora Vidová Hladká, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, Jan Votrubec: *Průvodce Českým akademickým korpusem 1.0.* Praha: Karolinum. 2007

Important publications from the last five years – papers

Hajičová, Eva – Cuřín, Jan – Hajič, Jan – Kučera, Ondřej – Vidová-Hladká, Barbora. *Jazyk a umělá inteligence: kudy a kam?* Praha: Academia, 2007.

Hajičová, Eva – Hladká, Barbora – Kučová, Lucie. An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In *Proceedings of the Workshop on Constraints in Discourse Structure Analysis.* 2006, pp. 82-89.

Hladká, Barbora – Kučera, Ondřej: A Corpus-based exercise book of Czech, In *Proceedings of the EUROCALL Conference (Mastering Multimedia: Teaching Languages Through Technology),* pp. 111, Coleraine, Northern Ireland, UK, 2007.

Hladká, Barbora – Králík, Jan. Proměny Českého akademického korpusu. *Slovo a slovesnost,* 67:179-194, 2006.

Ribarov, Kiril – Bémová, Alevtina – Hladká, Barbora. When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion., Prague Bulletin of Mathematical Linguistics, 2006, 1, 21-38.

Working experience in research teams

2004–present: PI of the „Resources and tools for information systems“ project (No. 1ET101120413) funded by GA AV ČR

2007-present: researcher at the „An electronic exercise book of Czech based on the Prague Dependency Treebank“ project (No. 207-10/257559) funded by GAUK

Scientific collaborators

Curriculum vitae: Jiří Mírovský

Born: April 20, 1973 in Kladno mirovsky@ufal.mff.cuni.cz
Address: Rabasova 1395, Slaný, 274 01

Affiliation:
Charles University in Prague, Faculty of Mathematics and Physics, Ke Karlovu 3, 200 02 Prague 2
Department: Institute of Formal and Applied Linguistics, Malostranské nám. 25, 118 00 Prague 1

Education:
1993–1998: Charles University in Prague, Faculty of Mathematics and Physics (computer science, Master degree)
2003–present: PhD studies at Charles University in Prague, Faculty of Mathematics and Physics (computational linguistics)

Professional experience:
2000–2004: Center for Computational Linguistics (Charles University in Prague)
2005–present: Institute of Formal and Applied Linguistics (Charles University in Prague)

Important publications from the last five years – books

Vidová Hladká, Barbora – Hajič, Jan – Hana, Jiří – Hlaváčová, Jaroslava – Mírovský, Jiří – Votrubec, Jan: *Czech Academic Corpus 1.0 - Guide.* Prague: Karolinum. 2007

Important publications from the last five years – papers

Mírovský, Jiří: Netgraph – Making Searching in Treebanks Easy, In: Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India, 8th - 10th January 2008, pp. 945-950.

Mírovský, Jiří: Towards a Simple and Full-Featured Treebank Query Language, In: Proceedings of ICGL 2008, Hong Kong, 9th - 11th January 2008, pp. 171-178.

Mírovský, Jiří – Panevová, Jarmila: Learning to Search in Prague Dependency Treebank, In: Proceedings of Grammar and Corpora 2007, Liblice, Czech Republic, 25th - 27th September 2007

Mírovský, Jiří: Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0, In: Proceedings of The Fifth International Treebanks and Linguistic Theories conference, Prague, Czech Republic, 1st and 2nd December 2006, pp. 211-222.

Smrž, Otakar – Pajas, Petr – Žabokrtský, Zdeněk – Hajič, Jan – Mírovský, Jiří – Němec, Petr: Learning to Use the Prague Arabic Dependency Treebank, In: Elabbas Benmamoun. Proceedings of Annual Symposium on Arabic Linguistics (ALS-19). Urbana, IL, USA, Apr. 1-3: John Benjamins, 2005.

Working experience in research teams
2000–2004: researcher at the project Center for Computational Linguistics (MŠMT LN00A063)
2002–2004: researcher at the international project MALACH (Multilingual Access to Large Spoken Archives, MŠMT 1P05ME786)
2004–present: researcher at the project "Data and Tools for Information Systems" (GA AV ČR 1ET101120413)

Curriculum vitae: Pavel Schlesinger

Date of Birth: February 6th 1979, Třinec schlesinger@ufal.mff.cuni.cz
Address: Habrová 300, 739 61 Třinec

Affiliation:
Faculty of Mathematics and Physics (MFF), Charles University in Prague (UK), Ke Karlovu 3, 121 16 Prague 2

Department: Institute of Formal and Applied Linguistics MFF UK, Malostranské nám. 25, 118 00 Prague 1

Education:

2005–present ongoing Ph.D. study in Computational Linguistics at MFF UK
1998–2005 Mgr. (equiv. level of M.S.) in Mathematical Statistics at MFF UK

Professional experience:

2004–present Institute of Formal and Applied Linguistics – machine learning, statistical methods in computational linguistics

2001-2002 Czech Statistical Office – processing and preparation of data and statistical methods

Publications - Journal articles and conference/workshop papers:

Pavel Pecina and Pavel Schlesinger: *Combining Association Measures for Collocation Extraction.* Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions, Sydney, Australia, July 2006.

Silvie Cinková, Pavel Pecina, Petr Podveský and Pavel Schlesinger: *Semi-automatic Building of Swedish Collocation Lexicon.* Proceedings of the fifth International conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, May 2006.

Jurjen Duintjer Tebbens and Pavel Schlesinger: *Efficient Implementation of Optimal Linear Discriminant Analysis.* Proceedings of the Seminar on Numerical Analysis (SNA'06), Sedlec-Prčice, Czech Republic, January 2006.

Jurjen Duintjer Tebbens and Pavel Schlesinger: *Improving implementation of linear discriminant analysis for the high dimension/small sample size problem.* Computational Statistics & Data Analysis, 52, 423 – 437, 2007.

Eduard Bejček, Pavel Straňák and Pavel Schlesinger. *Annotation of Multiword Expressions in the Prague Dependency Treebank.* Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), 793–798, India, January 2008.