Vernieuwingsimpuls / Innovational Research

Grant application form 2002

VENI scheme

Please refer to explanatory notes when completing this form

Registration form (basic details)

1. Details of applicant

Name, title(s): Bogdan BABYCH
Male/female: M
Birth date: 28.04.1974
Postal address: Oude Veurnestraat, 3, apt. 4 B, Ieper, 8900, Belgium
Telephone: +32 477 / 251 799
Fax: –
E-mail: babych@altern.org

2. Title of research proposal

Automatic approximation of formal grammars for speech and language processing systems

3. Summary of research proposal (max. 300 words, plus keywords)

Reusability of formal grammars and their adaptation for scalable systems are important problems in contemporary computational linguistics. A possible solution to these problems is to automatically create less complex approximations of full-scale grammars, which satisfy the requirements of specific applications architecture, its platform, and also give good performance on specific types of target text corpora.

Our research project focuses on investigating approaches to the development of automatic approximations of formal grammars, and on applying the approximations to the domain of text-to-speech technology, especially, for disambiguating homographs, determining location of accents, phrase boundaries of different strength, intonation types for phrases, as well as realisation of these prosodic properties in the speech signal.

In order to perform the research project the following ingredients are required: (1) a well-defined formal grammatical framework, (2) one or more large-scale grammars written in this framework, and (3) well-defined parsing algorithms for this framework. The XTAG English grammar, developed in the framework of Tree Adjoining Grammars (TAG), satisfies all these requirements, and will be used in the current research proposal, but other grammatical frameworks satisfying the minimal requirements will also be considered.

In our project we will derive 3 levels of approximation of the XTAG grammar with different computational complexity, (1) finite state approximations for part-of-speech taggers; (2) cascades of regular grammars that are dependent on the training corpus and target application; (3) full context-free or tree adjoining grammars restricted to subsets as determined by training corpus and statistics of rule applications. We plan to derive approximations with required properties out of stochastic TAGs, created for different types of corpora, and explore their reusability.

The goal of our research is to develop new approaches to the general problem of grammar approximation and create systematic and theoretically grounded techniques for this task.

Key words: Tree Adjoining Grammars; automatic approximation; computational complexity; reusability of grammars, scalable systems.

4. NWO Council area – GW

5. Host institution
Utrecht University, Utrecht institute of Linguistics OTS

Research proposal

Description of the proposed research (max. 2000 words)

AUTOMATIC APPROXIMATION OF FORMAL GRAMMARS FOR SPEECH AND LANGUAGE PROCESSING SYSTEMS

Research proposal

Bogdan Babych

a. Research topic. The development of formal grammars for morphosyntactic parsing of natural language is an important component of speech and language technology, and may considerably improve the quality of speech and language processing systems in the areas of machine translation, information extraction, question-answering, language generation, speech recognition and speech synthesis. Currently, several full-scale formal grammars, have been developed for different languages, for example: M-grammar for Dutch and English [Rosetta, 1994], [Odijk, 1993], English Resource Grammar (based on HPSG and Construction Grammar frameworks) [Pollard, Sag, 1994], [Fillmore, Kay, 1993] and XTAG English Grammar, developed in the format of Tree Adjoining Grammars (TAGs) [Joshi, Schabes, 1997], [Doran et al., 2000]. These grammars formalise valuable linguistic knowledge about morphosyntax, and can be treated as application-independent syntactic databases. *Reusability* of formal grammars (when a single grammar is used for multiple NLP and speech technologies) and adaptation of the grammars for *scalable systems* (the systems that can be automatically generated from a single base system for platforms of different sizes) are important problems in contemporary computational linguistics. This is due to the fact that full-scale formal grammars are often not directly usable in actual applications for a variety of reasons:

1. they are inherently too complex (in terms of computational complexity);
2. they always require additional modelling of world knowledge and situational knowledge to perform adequately; otherwise there will inevitably be too many ambiguities, which will only increase the more sophisticated the lexicon and the grammar becomes;
3. the targeted platform may impose additional restrictions in terms of processing power and memory, making it impossible or difficult to use full-scale grammars directly;
4. however large the grammar and the lexicon may be, full coverage of the language is still far away for all grammars.

As a result, formal grammars are often built from scratch for each new task or application, which is undesirable since it leads to duplication of effort, mutual inconsistency, more difficult maintenance, etc.

A possible solution for this problem is to automatically create less complex approximations of full-scale grammars, which satisfy the requirements of specific applications architecture, its platform, and also gives good performance on specific types of target text corpora for such applications. There were reports in the literature about practical experiments with automatic approximation and optimisation of formal languages (in particular, regular approximation of context free languages [Nederhof, 2000], determinizing finite-state automata (FSA) [van Noord, 2000]). But still, there remains a problem of automatic approximation of real full-scale formal grammars to different types of less computationally expensive formalisms.

Our research project focuses on investigating approaches to the development of automatic approximation of formal grammars, and on applying the approximations to the domain of text-to-speech technology, especially, for disambiguating homographs, determining location of accents, phrase boundaries of different strength [Theune et al., 1997], intonation types for phrases [Collier, 't Hart, 1972], [Willems at al., 1988], [Bryzgunova, 1969], as well as realisation of these prosodic

properties in the speech signal (pitch movement patterns, melodic highlighting of syntactic boundaries, segment lengthening, pausing).

In order to perform the research project a well-defined formal grammatical framework, one or more large-scale grammars written in this framework, and well-defined parsing algorithms for this framework are required. The XTAG system, mentioned above, satisfies all these requirements, and will be used in the current research proposal, but other grammatical frameworks satisfying the minimal requirements – if available – will also be considered.

The XTAG framework is especially interesting for a variety of reasons:

- it is a well-defined grammatical framework; the framework is known to be able to deal with properties that take natural language beyond the weak generative capacity of (CFGs), such as crossing dependencies in Dutch and Swiss German;
- a parsing algorithm has been defined for it, and its time and space complexity properties are well-known;
- a large grammar for English has been written in this framework and grammars for other languages are being developed;
- the code sources of the grammar and parser are available and open, and currently still maintained by the original developers [Abeillé, Rambow, 2000].

We will focus on the problem of automatic finite-state approximation of TAGs for predicting prosodic properties, needed for text-to-speech (TTS) [Collier, Landsbergen. 1995], [Vronis et al., 1997] and data-to-speech systems [Theune et al., 1997; 2000]. It has been shown that finite-state prosodic models are appropriate for TTS applications [Abney, 1995, p. 3-7], [Maireüil, d'Alessandro, 1998, p. 2-3], [Bondarko, 2000, p. 124-127], [Fitzpatrick, 2001, p.549]. For example, cascades of FSA usually perform phrasing in TTS systems, where lower-stratum automata group lexical items into chunks, which, in their turn, are grouped by higher-stratum automata. The number of strata is always limited, so partial parsing for TTS can be done with a fixed amount of memory, in linear time and without parsing ambiguities. Despite possible inaccuracies of this approach (e.g., difficulties with recognising cases of embedding higher-level chunks into lower level ones), its general performance for intended types of input texts is good. Parsing with complete XTAG system is redundant for TTS tasks, mainly because of high degree of ambiguity.

In our research project the following problems will be investigated:

1. Deriving 3 levels of approximation of the XTAG grammar with different computational complexity, creating a spectrum of its scalability:

- Building a *finite state approximation* for part-of-speech (PoS) taggers and investigating its efficiency for the tasks of homograph disambiguation and text normalisation, needed for TTS;
- Building a phrase parser ("chunker") using a *cascade of regular grammars* (stochastic or deterministic) that is dependent on the training corpus and target application. Investigating the efficiency of the phrase parser for predicting prosodic properties of sentences for TTS systems;
- Building full *context-free or tree adjoining grammars* restricted to subsets as determined by training corpus and statistics of rule applications. Investigating the performance of these grammars for the TTS tasks of homograph disambiguation and prediction of sentence prosodic properties, as compared to the performance of the two lower level models: the finite state PoS taggers and the cascaded finite-state phrase parser.

We will try to develop new approaches to the general problem of grammar approximation, based on our work on the TAG scalability. The goal of the research is to create systematic and theoretically grounded techniques for automatic approximation of formal grammars.

2. Investigating properties of approximations created on different types of corpora. The XTAG system is supposed to cover morphosyntactic structures represented in all functional styles of English, but its efficient approximations have to focus on particular types of corpora, where certain collocations of elementary trees have varying productivity. In our approach different types of corpora

are expected to result in different approximations. An important problem to be investigated is how large and syntactically diverse the training corpus has to be, in order to produce efficient approximation for texts of a particular type. We are going to address this question by building the approximations for different types of corpora, e.g., medical texts and judicial texts. We will also build experimental approximations on a corpus of fiction prose, which is expected to have richer vocabulary and more diverse inventory of syntactic structures. This investigation will allow determining the optimal size of a training corpus as a function of formal measures of stylistic diversity in its lexicon and syntax [Martynenko, 1996].

3. Investigating reusability of XTAG approximations. Besides the domains of TTS and data-to-speech technology, automatic grammar approximations can be used in other areas. E.g., though for current automatic speech recognition (ASR) systems usually simple (finite state or CFG) grammars are used to characterise finite languages, the development of conversational speech recognition systems will require more complex and open-ended grammars characterising infinite languages [Chelba, Jelinek, 1998], [van Noord et al., 1999]. For these purposes high-level approximations of TAGs can be used, e.g., to stochastic CFGs [Caroll, Weir, 1997] or deterministic pushdown automata (PDA) [Partee, 1993, p. 488].

Another problem is automatic generation of synchronous TAGs. Since the development of isomorphic M-grammars proved to be very successful for compositional machine translation (MT) [Rosetta, 1994], [Odijk, 1993], there have been suggestions to use similar formalism of synchronous TAGs for purely surface-based MT [Abeillé et al., 1990], but yet no real system has been developed in the TAG framework. One of the reasons is that manual encoding of synchronous TAGs is a very large and complex task. It has to be investigated if synchronous TAGs can be derived automatically from parallel corpora, e.g., from an aligned treebank, and how large such treebank has to be to ensure acceptable quality.

These applications set additional requirements on formal grammars. In order to explore reusability of the XTAG system, we will address the problem how TAG approximations can meet these demands.

4. Investigating theoretical implications of the grammar approximation techniques for complexity measures of language. The problem of natural language complexity of still is an open issue in computational linguistics. Approximations of TAGs, build on real corpora, will provide valuable statistics about how frequently natural language goes beyond the generative capacity of regular grammars, deterministic PDA, CFGs, or even TAGs (the examples of scrambling, which require multi-component TAGs [Weir, 1998]). In this respect the question 'if the natural language is regular or context free', can be reformulated as 'to which extent it is regular or context free'. We expect to find out exact values for syntactic complexity and diversity in different types of corpora.

In our opinion, the following aspects of our research are innovative:

- Approximating formal grammars for real speech and language applications is perhaps not fully new, but our attempt to derive a whole range of approximations of different scale in a systematic manner is innovative;
- Using TAG grammars (and the XTAG system in particular) to derive approximations is original;
- Our attempts to apply grammar approximation techniques, known for other grammar types, to a new grammar type and to develop complete new approximation techniques – are innovative.

b. Approach. Our approach consists of building stochastic TAGs and deriving approximation with the required properties out of them. We suggest building approximation of formal grammars using automatically created stochastic TAGs. [Joshi A., 1999]. Frequency information can be encoded in different ways in lexicalized grammars [Caroll, Weir, 1997]. We will adopt the most powerful version

of stochastic TAGs, based on globally-dependent frequencies – an approach developed in the framework of Data-Oriented Parsing [Bod, 1998]. A method of extraction of stochastic lexicalized trees from existing treebanks has been proposed in [Neumann, 1998], [Bod, 1999]. In our project we cannot use this method directly, because we plan to build corpus-dependent approximations, so we would need to develop a clean treebank for each type of target corpora, (which requires much time, and so diminishes the value of grammar approximation as a purely automatic procedure).

Instead, we propose a method of building stochastic TAGs out of complete sets of non-ranked ambiguous parses, produced by the current XTAG system. We suggest exploiting varying degree of ambiguity for the same elementary trees in parsed corpora. The idea is that most probable elementary trees will occur both in ambiguous and in unambiguous (or less ambiguous) positions. We will determine these probabilities by processing ambiguous sets of parses in corpora and creating statistical combinatory table for adjunction and substitution nodes in each elementary tree, found in the corpus. The most probable collocations of elementary trees are good candidates for being interpreted as 'terms' or 'entities' in the subject domain of the training text corpora. Our approach in this respect resembles the classical N-gram model, but it is build from ambiguous parses in corpora, rather then from linear sequences of words.

We will use frequency information in the stochastic TAGs to build approximations with different properties. For example, for TTS finite-state approximations we will determine (a) the hierarchy of FSA that check PoS codes of words. The structure of this hierarchy should give the optimal coverage of the corpus; (b) the most frequent exceptions to the hierarchy that will be merged by highest-priority-FSA, which will check lexical items, instead of PoS codes. In terms of prosodic features such frequent structural collocations will have less chance to be separated by a prosodic boundary (which is normally expected for terms and entity names). Similar techniques will be developed to create other types of TAG approximations.

c. Plan of work. We suggest the following plan for carrying out the research project:

August 2002 – August 2003:

Developing stochastic TAGs for different types of corpora

- Reading literature, installing XTAG system, familiarizing oneself with this system, developing algorithms for stochastic processing of TAGs;
- Purchasing legal and medical corpora from ELRA, LDC, or other organizations;
- Downloading freely available corpora of fiction texts from the Internet;
- Developing programs for producing stochastic TAGs;

August 2003 – February 2004:

Creating finite-state approximations of TAGs for TTS technology

- Deriving finite-state approximations of TAGs for PoS taggers;
- Implementing a PoS tagger for homograph disambiguation tasks in TTS applications;
- Evaluating of the quality of the finite-state approximations for different types of corpora;
- Working at the University of Pennsylvania with the group of Prof. A. Joshi on XTAG system for 3-4 months;

February 2004 – August 2004:

Creating cascaded regular approximations of TAGs for prosody generation

- Deriving cascaded regular grammars from stochastic TAGs, generated from different types of corpora;
- Implementing a phrase parser for predicting phrase boundaries in TTS systems, based on the cascaded regular approximations;

- Extending the phrase parser with additional features for determining location of accents and intonation types of phrases;

August 2004 – February 2005:

Creating restricted CFG and TAG grammars for TTS applications

- Deriving stochastic CFG approximations and deterministic PDA approximations from stochastic TAGs ;
- Deriving restricted deterministic and stochastic TAGs, optimised for different types of corpora;
- Evaluating the performance of the restricted grammars, when these grammars replace finite-state and cascaded regular approximations in the PoS tagger and the phrase parser.

February 2005 – August 2005:

Investigating XTAG reusability and measures of syntactic complexity and diversity of corpora, used in the approximation techniques

- Exploring application-specific requirements for formal grammar approximations in other domains of language and speech technology, such as ASR and MT; adapting the approximation algorithms for meeting these requirements;
- Measuring distribution of syntactic structures of different complexity in corpora; investigating the values of syntactic complexity and diversity in various types of texts;
- Estimating optimal size and syntactic diversity of a training corpus needed to create XTAG approximations for different types of applications.

Local, national and international collaboration. The research will be carried out in the Computational Linguistics and Logic group of UIL OTS, in which prof.dr. Michael Moortgat, prof.dr. Jan van Eijck (UiL OTS/CWI, Amsterdam) and prof.dr. Jan Odijk (UiL OTS/Scansoft) are operative. This group is involved in the syntactic annotation and prosodic annotation projects (Van der Wouden, Hoekstra, Goddijn) of the NWO-programme Corpus Spoken Dutch (CGN). The proposed research project will be able to benifit from the experience acquited in these projects.

Extensive expertise on intonation, prosody is availbale in the Phonology & Morphology group (Dr. Kager, Prof.dr. Zonneveld) and Phonetics group (Prof.dr. Nooteboom, Dr. H. Quené). UiL OTS has a long standing expertise this area through the involvement of Prof.Ir. Landsbergen (diss. of Sima'an and Huijsen, among others).

Bibliography:

[Abeillé et al., 1990] – Abeill'e, A., Y. Schabes, and A. K. Joshi (1990) "*Using Lexicalized Tags for Machine Translation*," in Proceedings of the International Conference on Computational Linguistics (COLING '90), Helsinki, Finland.

[Abeillé, Rambow, 2000] – Abeillé, Rambow. Tree Adjoining Grammars. An overview. In.: Tree Adjoining Grammars, 2000, p. 19.

[Abney, 1995] – Abney, Steven. 1995. *Chunks and dependencies: Bringing processing evidence to bear on syntax*. In Computational Linguistics and the Foundations of Linguistic Theory. CSLI.

[Bod, 1998] – Bod, R. 1998. Beyond Grammar: An Experience-Based Theory of Language. CSLI Publications /Cambridge University Press.

[Bod, 1999] – Rens Bod. 1999. Extracting stochastic grammars from treebanks. In Journées ATALA sur les Corpus annotées pour la syntaxe. Talana, Paris VII.

[Bondarko, 2000] – Bondarko, ed. Phonology of Speaking. (Бондарко Л.В., ред. Фонология речевой деятельности)., 2000, p. 124-127 (in Russian).

[Bryzgunova, 1969] – Bryzgunova E.A. Sounds and intonation of Russian speech (Брызгунова Е.А. Звуки и интонация русской речи), Moscow, 1969 (in Russian).

[Caroll, Weir, 1997] – Caroll, John and David Weir. 1997. Encoding frequency information in lexicalized grammars. In ACL/SIGPARSE workshop on Parsing Technologies, MIT, Cambridge.

[Chelba, Jelinek, 1998] – Ciprian Chelba and Frederick Jelinek. Exploiting syntactic structure for language modeling. In

Proceedings of COLING-ACL,volume 1, pages 225--231. Montreal, Canada, 1998.
[Collier, Landsbergen. 1995] – Collier, R. and J. Landsbergen. 1995. Language and speech generation. Philips Journal of Research, 49(4):419--437
[Collier, t'Hart, 1972] – Collier, R. and Hart, J.'t Perceptual experiments on Dutch intonation. In A.Rigault and R.Charboneau (eds), Proceedings of the 7th international Congress of Phonetic Sciences (pp. 880-884), The Hague, Paris: Mouton.
[Doran et al., 2000] – Doran et al. Evolution of the XTAG system. In: Tree Adjining Grammars, 2000, .p. 371-404
[Fillmore, Kay, 1993] – Fillmore, Charles and Paul Kay. 1993. Construction Grammar. University of California, Berkeley. Language 64: 501-538.
[Fitzpatrick, 2001] – Fitzpatrick, Short report: The prosodic phrasing of clause-final prepositional phrases., Language, No. 3, 2001, p. 544-561.
[Joshi, 1999] – Joshi A. (1999) Explorations of a domain of locality. CLIN'99. Utrecht.
[Joshi, Schabes, 1997] – Joshi, A. K. and Y. Schabes. 1997. *Tree-adjoining grammars*. In Grzegorz Rozenberg and Arto Salomaa, editors, Handbook of Formal Languages. Vol 3: Beyond Words, chapter 2, pages 69-123. Springer-Verlag, Berlin/Heidelberg/New York.
[Maireüil, d'Alessandro, 1998] – Maireüil, d'Alessandro. 1998, Text Chunking for Prosodic Phrasing in French. In: Proc. 3rd ESCA Workshop on Speech Synthesis - 98
[Martynenko, 1996] – Martynenko G. Ya. Complexity of syntactic structures and stylistic diagnostics. In: Applied Linguistics. Ed. A.S. Gerd, (Прикладная лингвистика. Ред. С.Я Герд) S. -Petersburg, 1996 (in Russian)
[Nederhof, 2000] – M.-J. Nederhof, 2000, Practical experiments with regular approximation of context-free languages. In: Computational Linguistics, 26 (1), p. 17-44.
[Neumann, 1998] – Neumann, G. 1998. "Automatic Extraction of Stochastic Lexicalized Tree Grammars from Treebanks", Proceedings of the 4th Workshop on Tree-Adjoining Grammars and Related Frameworks , Philadelphia, PA.
[Odijk, 1993] – Jan Odijk. *Compositionality and Syntactic Generalizations*. PhD thesis, University of Tilburg, 1993.
[Pollard, Sag, 1994] – Pollard, Carl and Ivan A. Sag. 1994. Head-Driven Phrase Structure Grammar. Chicago: University of Chicago Press and Stanford: CSLI Publications.
[Rosetta, 1994] – M.T. Rosetta, editor, *Compositional Translation*. Kluwer Academic Publishers, Dordrecht, 1994.
[Theune et al., 1997] – Theune, M., Klabbers, E., Odijk, J., and de Pijper, J.R., "Computing Prosodic Properties in a Data-to-Speech System," Workshop on Concept-to-Speech Generation Systems, (E)ACL, Madrid, 39-46, 1997
[Theune et al., 2000] – Theune, M., E. Klabbers, J.R de Pijper, E. Krahmer, en J. Odijk. From Data to Speech: A General Approach, *Natural Language Engineering*, 7(1), pp. 1-40, 2000
[van Noord et al., 1999] – Gertjan van Noord, Gosse Bouma, Rob Koeling, and Mark-Jan Nederhof. Robust grammatical analysis for spoken dialogue systems. Journal of Natural Language Engineering, 1, 1999.
[van Noord, 2000] – G. van Noord, 2000, Treatment of epsilon moves in Subset Constructions. In: Computational Linguistics, 26 (1), p. 61-76
[Vronis et al., 1997] – Vronis, J., Di Cristo, P., Courtois, F. & Lagrue, B. 1997. A stochastic model of intonation for text-to-speech synthesis. Proceedings Eurospeech '97 (Rhodes) 5: 2643-2646.
[Weir, 1998] – Characterizing mildly context sensitive grammar formalisms. Doctoral dissertation, Department of Computer and Information Science, University of Pennsylvania.
[Willems at al., 1988] – Willems , M.J., Hart, J.'t and Collier, R., 1998, English intonation from a Dutch point of view. Journal of the acoustic society of America, 84, 1250-61.

Cost estimates

Staff costs (per year, in fte, tenured/fixed term, see explanatory notes)

	gross	+70%	total
Jaar 1 (11.1)	€ 41.412	€ 28.989	€ 70.401
Jaar 2 (11.2)	€ 44.256	€ 30.979	€ 75.235
Jaar 3 (11.3)	€ 47.244	€ 33.071	€ 80.315
	€ 132.912	€ 93.039	€ 225.951

Personnel costs are based on information of the bureau of the Faculty of Arts of Utrecht University

Non-staff costs (per year, see explanatory notes)

Jaar 1	Travel costs:		
	• Visiting conferences	€ 3.500	
	Purchasing equipment:		
	• A dedicated computer with a large hard disk (> 80 Gb) for processing corpora, CDRW; software	€ 5.000	
	Purchasing linguistic resources:		
	• legal and medical corpora from ELRA or LDC	€ 5.000	
		Subtotal for the year 1	€ 13.500
Jaar 2	Travel costs:		
	• Visiting conferences	€ 3.500	
	• Working in XTAG group of Prof. A.Joshi at the University of Pennsylvania, USA, 3-4 months, € 2000 per month	€ 7.000	
		Subtotal for the year 2	€ 10.500
Jaar 3	Travel costs:		
	• Visiting conferences	€ 3.500	
		Subtotal for the year 3	€ 3.500
		Total:	€ 27.500

The budget is approved by the the director of the Utrecht institute of Linguistics OTS.

Curriculum vitae

-Personal details

Title(s), initial(s), first name, surname:	Bogdan BABYCH
Male/female:	M
Date and place of birth:	28.04.1974, Kirovograd, Ukraine
Nationality:	Ukrainian

-Master's ('Doctoraal')

University/College of Higher Education:	Kyiv Taras Shevchenko University
Date:	28.06.1996
Main subject:	Ukrainian Philology and Computational Linguistics

-Doctorate

University/College of Higher Education:	National Ukrainian Academy of Sciences Institute of Linguistics / Ukrainian Language Information Fund
Date:	27.06.2000
Supervisor ('Promotor'):	Prof. Dr. Volodymyr V. CHUMAK
Title of thesis:	Interpretation model of surface syntactic structures in Ukrainian

-Work experience since graduating
(per appointment: fte, tenured/fixed-term, see notes)

August 2000 – December 2001: Lernout & Hauspie Speech Products NV, Ieper, Belgium
Corporate R&D, Linguistic Engineering Department, Text to Speech division
Computational Linguist

February 1999 – July 2000: Language Information Fund
of the Ukrainian National Academy of Sciences, Kyiv, Ukraine
Research fellow

-Brief summary of research over last five years (max. 250 words)

In 1996–1999, at a post-graduate programme, I developed a word order variation model for Ukrainian, in the framework of Tree-Adjoining Grammars. The model predicts a set of all possible synonymous sentences with varying word order for a given structural organisation of constituents. Differences in distribution were found for configurational and unconfigurational word order (which appears in all styles) on the one hand, and unprojective word order (which systematically appears only in spontaneous speech and poetry) on the other.

I applied the model to investigating restrictions on centre embedding in Ukrainian, and suggested that the acceptability changes gradually, until four levels of embedding are reached. Examples of unacceptable relative clause embedding with fewer levels were shown to be ungrammatical, violating binding requirements. This data suggests that the processing complexity of syntactic structures correlates with their distribution and perception difficulties, and that computationally redundant syntactic theories are psychologically less plausible.

Improvements for a deep syntactic representation format were proposed for modelling Slavic word order variation, the format was also used for representing structures of semantic primitives; and on this basis a technique was developed of automatically deriving formal semantic representations

from word definitions in monolingual dictionaries for large-scale NLP systems.

In 2000–2001, at "L&H Speech products" I proposed solutions to several text processing problems for Ukrainian and Russian text-to-speech systems, e.g., I developed data-driven algorithms for part-of-speech tagging and stressing, a finite-state parsing algorithm for predicting phrase boundaries in Russian clauses, morphological analysis and generation modules for text normalisation.

-International activities:

Summer Schools:

July, 2001 Netherlands Graduate School of Linguistics (LOT),
Utrecht University, The Netherlands

1997 (June – August) Linguistic Institute of the Linguistic Society of America
"Languages in Linguistics",
Cornell University, Ithaca, New York, USA

EU project TEMPUS:

January 1998 – February 1998: University of Granada, Spain
Working on multilingual dictionaries for "TEMPUS translation tools" CD
within the joint Spanish-Ukrainian lexicographical project TEMPUS
of Kyiv University and the University of Granada, sponsored by the EU

-Other academic activities

Teaching:

September 1998 – May 1999 – teaching assistant
at the Department of Foreign Philology of Kyiv University
course title: "Computer Aided Translation"

Presentations on conferences:

January 1999 – All-Ukrainian conference
"Semantics, syntactics and pragmatics of Speaking", Lviv University, Lviv, Ukraine

May 1998 – International conference "Computational Linguistics and Teaching Foreign Languages", Lviv Technical University, Lviv, Ukraine.

December 1997 – International conference on Ukrainian spelling reform
Kyiv-Mohyla Academy, Kyiv, Ukraine.

-Scholarships and prizes

- Scholarship of the Ukrainian National Academy of Sciences for young scholars.
- Tuition grant from the Linguistic Society of America for 1997 LSA Linguistic Institute
- Diploma with Honours from Kyiv Taras Shevchenko University
- Personal Scholarship from Kyiv Taras Shevchenko University
- Scholarship of the Students Scientific Society of Kyiv University

List of publications

-National (refereed) journals

1. Systems of syntaxeme groups and their procedural semantics. In: “Movoznavstvo” ("Linguistics", the scholarly theoretical journal of the O.O.Potebnya Institute of Linguistics and the Institute of Ukrainian), 1998.—№6.—Pp. 55-62. – **(S)(!)**
2. Role of pragmatic context for disambiguating syntactic structures. In: "Lingual and Conceptual Models of the World", Kyiv, Kyiv Taras Shevchenko University, 1998, Pp. 25-30. – **(S)**
3. Representing and interpreting the ambiguous deep structures. In: “Ukrajins’ke movoznavstvo” ("Ukrainian Linguistics", Kyiv Taras Shevchenko University), 1997.— Vol 21.—Pp. 89-100. – **(S)**
4. Diphthongs in the Northern Ukrainian dialects and in the history of the Ukrainian Language. In: “Visnyk Kyivs’koho Universytetu imeni Tarasa Shevchenka” (Bulletin of Kyiv Taras Shevchenko University), 1994, Vol 2.—Pp. 99-102.

-Other

1. Lexical Semantics in the syntactic structure of a text: formal representation and interpretation. In: Proceedings of all-Ukrainian conference “Semantics, syntactics and pragmatics of speech”, Lviv, Ukraine, January 1999.—P.101-108 – **(S)(!)**
2. Correcting grammatical inconsistencies of deep syntactic constructions in automatic grammar control systems. In: Proceedings of International conference on Ukrainian spelling reform, Kyiv, Ukraine, 1997.– P.22-23

Vernieuwingsimpuls / Innovational Research

Grant application form 2002

VENI scheme

Please refer to explanatory notes when completing this form

Post to NWO

To streamline the processing of applications, please complete the form below and post a print-out of this page together with the relevant documents to NWO.

I the undersigned declare that I have today posted (tick relevant documents):

☑ **Reprints of two main publications**
(obligatory for all applicants)

Official declaration that my thesis manuscript has been approved and a date fixed for its defence
(obligatory for applicants for VENI grants who have not yet received their doctorates)

Institutional guarantee from Board ('Inbeddingsgarantie College van Bestuur')
(optional for VENI)

Address list of 'non-referees'
(optional for all applicants, max. 3, see explanatory notes)

Name of applicant: Bogdan BABYCH

Place: Ieper, Belgium

Date: 12.02.2002

Postal address: Oude Veurnestraat, 3, apt, 4 B, Ieper, 8900 Belgium

NWO Council area: ALW / CW / EW / GW / MaGW / MW / TW / Other

---

The documents should be received by NWO before the deadline for submissions. Send them to:
NWO
Dr W.A. van Donselaar
P.O. Box 93138
2509 AC The Hague
(The Netherlands)