



Department of Computational Linguistics
Annual report 2004
January 2005

Content

1	Staff	2
2	Research Funds	3
3	History	3
4	Annual report 2004	4
4.1	Research projects in 2004	5
4.1.1	EU projects	6
4.1.2	Projects with funding from other external resources	6
4.1.3	Research projects in collaboration with The Danish Centre for Terminology	8
4.1.4	Projects with funding from internal resources	9
4.1.5	Other research projects	13
4.1.6	PhD Projects	14
4.2	Other academic activities in 2004	14
4.2.1	PhD study programme	14
4.2.2	Network activities	15
4.2.3	International conferences	15
4.2.4	Participation in interdisciplinary work groups and committees	15
4.2.5	Participation in collaborative projects	16
4.2.6	Evaluation and reviews	16
4.2.7	Publications	16
5	Teaching activities	17
5.1	The Master's programme in Computational Linguistics	17
5.2	MA supplementary course in Language Policy	17
5.3	The BA programme in Language Technology and English	17
5.4	BA in Management of Information, Communication and Knowledge (MICK)	18
5.5	The Master of Language Administration programme	18
5.6	Initiatives in further education	18

CBS Faculty of Languages, Communication and Cultural Studies
Department of Computational Linguistics

Bernhard Bangs Allé 17B
2000 Frederiksberg
Telephone: +45 38 15 31 36
Fax: 38 15 38 20
Email: ln.id@cbs.dk
<http://www.id.cbs.dk>

1 Staff

Head of Department

Sabine Kirchmeier-Andersen was head of department in the period covered by the report.

Executive Committee

Sabine Kirchmeier-Andersen, Hanne Erdman Thomsen, Steffen Leo Hansen, Lene Nissen

Research Staff

Full Professor:

Bodil Nistrup Madsen
Per Anker Jensen

Associate Prof.:

Stefen Leo Hansen
Sabine Kirchmeier-Andersen
Bjarne Ørsnes
Hanne Erdman Thomsen
Henrik Selsøe-Sørensen (50%, the rest at FIRST)
Daniel Hardt
Peter Juel Henrichsen

Assistant Research Prof.:

Matthias Trautner Kromann
Peter Rossen Skadhauge

System administrator:
Chief Developer:

Kåre Hviid
Bo Krantz Simonsen

PhD Students:

Nina Frederiksen (maternity leave untill August 15th)
Tina Nielsen (50% untill June 1st)
Ekaterina Mhaanna (maternity leave)
Lone Bo Sisseck
Jakob Halskov

Others:

Prof. Carl Vikner has since his retirement been affiliated with the department as external PhD supervisor and participant in the OntoQuery and Caos projects.

Administrative staff:

Academic officer:

Stig W. Jørgensen (50%)
Birgitte Laursen (from August 1st)

Adm. officers:

Lene Nissen (32 hours) (untill July 1st)
Gitte Jørgensen (25 hours)
Merete Ørslev Christensen

Adm. assistant:

Rie Astrup (25 hours)

2 Research Funds

Basic funding 4,8 full-time equivalents (FTE) (2,8 + 2 ph.d.)

Department of Computational Linguistics 2004 funding allocation			
		2003	2004
Funding from external resources			
Danish Research Agency (SHF) CMOL	Release/Loan/Research ass.		1.481.380
Danish Research Agency (Language Technology)	Network activities		30.000
Danish Research Agency (STVF) (SDMT)	Ph.d. grant + research ass.		117.500
Danish Research Agency (STVF) (OntoQuery)	Ph.d. grant (12 mdr).		300.000
Danish Research Agency (STVF) (FOVITS)	Network + course development		7.500
Total	FTE 4,84	1.096.000	1.936.380
Funding from internal resources			
Faculty	Promotion of Language Technology		60.000
Senior staff salary allocation	Terminology database project		821.450
	Professor salary, research part		112.500
	Communication and applications		200.000
Departmental research funds	IT Focus development		1.075.000
Total	FTE 5,67	2.203.000	2.268.950
Forskningsårsværk i alt		FTE 10,51	3.299.000
			4.205.330

The table shows that external funding has almost doubled during 2004 and internal funding has increased by 10 %. In total the research and development funds of the department have increased by almost 1 mill. kroner corresponding to an increase by 27%.

3 History

The Department of Computational Linguistics was established June 1st 1985 with the purpose of developing the academic basis for a computational linguistics programme, planning and teaching the courses of this programme, and doing research in computational linguistics, particularly within the following areas: Formal syntactic and semantic analysis of language for specific purposes (LSP), modelling and representation of knowledge relevant to LSP, natural-language interfaces, automatic translation, and computational terminology & lexicology. Within these areas, the main focus is on computational linguistic issues that involve Danish.

The department's research staff are active in inter-departmental and inter-institutional research projects, and for a number of years the department has worked in close collaboration with the Danish Centre for Terminology (*DANTERMcentret*) on presenting the department's research results in shared projects with a

number of Danish companies. In 2004 it was decided that the Danish Centre for Terminology will continue as a research and development centre of the department as from March 2005.

The department is responsible for courses at graduate level in computational linguistics (CLM), at undergraduate level in IT and English, at professional master's level in Master of Language Administration (MLA), and for single courses at the BA programme in business language. In addition, the department holds PhD courses, as well as graduate and undergraduate courses in the languages programmes, and Open University. The department issues a line of publications titled LAMBDA.

4 Annual report 2004

The department has become the leading centre for formal research in language (with main emphasis on Danish), where the results are used in NLP systems. We are now striving for the following goals:

- The department must maintain this position
- The research in the department must meet international standards

Criteria for success: At least in three core areas research must meet international standards, i.e. there must be international refereed publications, the department should host international conferences and network meetings, members of the department should be invited speakers at international conferences, and research results should be part of Scandinavian or international standards.

Results: The department has produced international refereed publications in the following areas: Spoken language analysis, ontology structuring and statistical processing. Preparations have been made to host the international conference TKE (Terminology and Knowledge Engineering in August 2005), international network meetings on Computer-assisted language learning and discourse processing. The external research funding of the department has increased by 27% and the department has been a partner in a number of collaborative Nordic and EU applications.

- The department must develop in close co-operation with the private and public sector

Criteria for success: The department must be the preferred partner for companies when it comes to consultancy, development of language technology and courses on language technology. At least two companies should choose to enter into projects such as joint corporate PhD projects, hosting students projects or research projects.

Results: The department has engaged in bilateral agreements with Mikroværkstedet A/S about student projects and research cooperation, and with IBM about sponsorship of hardware and software as well as research cooperation. Furthermore, contacts have been made with Ankiro in order to prepare a corporate PhD project in the area of ontology structuring. The close cooperation with companies and organisations associated with the DANTERMcentre has been continued and extended. Furthermore, the department has participated in several expert committees of the ministry of science, technology and development and the national research council providing input and guidelines for government funding in the area of language technology and computational linguistics.

- The department offers study programmes and courses in language technology and computational linguistics at all levels and attracts new students.

Criteria for success: The programmes offered by the department must be the preferred choice of students, when it comes to language technology and computational linguistics. The number of students on the present programmes must be increased. New programmes and courses must be developed.

Results: The number of students in the BA and regular master courses (cand.ling.merc.) has unfortunately decreased further in 2004, whereas the number of students in the vocational master (MLA) is increasing. Since internal funding for the development of new programmes has not been available, the department has participated in several applications for course development.

The department is developing a new BA in language technology which is to be integrated into the BA in international business communication offered at the Faculty of Languages, Communication and Cultural studies.

The department is participating in the development of an international BA in Management of Information, Communication and Knowledge in close cooperation with several departments from both faculties at CBS.

The department has developed 4 courses for high school teachers in language technology in order to promote language technology for students at an early stage. One course was established with good results. However, the resources needed to promote further courses were not available.

The department was granted funding via the EU Erasmus Mundus programme for the promotion of higher education abroad. The funding is used to promote the MLA programme world wide in English.

Together with the universities in Saarbrücken, Prag, Bolzano, Nancy1+2 and Malta the department is preparing another application for Erasmus Mundus for the development of a european master in language technology.

The department is participating in the FOVITS project (Forskningsbaseret videreuddannelse i Sprogteknologi – Research-based Further Education in Language Technology). Together with the universities in Copenhagen (CST) and Ålborg, short courses are developed primarily for companies.

- The department contributes to spreading the general knowledge of computational linguistics and language technology and attracting public focus and funding to the area.

Criteria for success: The department must be visible in the public debate on language technology through popular articles and focus in the media on language technology. The web pages of the department must be revised and supplemented with an online show-room for language technology. The different projects must focus more on the user interfaces to their programs in order to make them more understandable and approachable in demonstration situations.

Results: During 2004 the department has been more visible than ever before in the public debate. Researchers have participated in TV-shows, Radio interviews, public conferences, feature articles in leading newspapers and popular articles in various contexts.

A special website was created introducing language technology to students with links to various applications that the students could try out for themselves (www.id.cbs.dk/sprogtek).

The department received internal funding to employ a researcher devoted to communication issues and promotions as well as EU and Nordic applications.

The researchers associated with the DANTERMcentre have made particular efforts to promote language technology in courses for companies and in direct contact with companies.

The department has maintained close contact with ITEK (Dansk Industri) and has become a member of DEAs (Dansk Erhvervsakademi) Network for language technology.

4.1 Research projects in 2004

Below, the department's research projects are described. Extended descriptions are given for new projects, whereas research reports from previous years should be consulted for full descriptions of ongoing projects.

4.1.1 EU projects

4.1.1.1 LATER – Language Technology Erasmus Mundus Programme.

LATER focuses on promoting world-wide European higher education in the fields of Computational Linguistics and Language Technology via the master programmes of the university departments which participate in the "LATER" consortium: the Department of Computational Linguistics and Phonetics (CoLi) of Saarland University (co-ordinator), the Institute of Formal and Applied Linguistics (UFAL) at the Faculty of Mathematics and Physics of the Charles University in Prague, the Department of Computer Science and Artificial Intelligence of the University of Malta, and the Department of Computational Linguistics of the Faculty of Modern Languages at the Copenhagen Business School. Specifically, "LATER" focuses on making European higher education in Computational Linguistics and Language Technology attractive to students and scholars from third countries. To this end, and in order to enable dissemination of master programmes and courses in Computational Linguistics and Language Technology and attract third-country students, "LATER" provides for the organisation of workshops and conferences, as well as the development of distance learning tools, and distance education modules. The grant for LATER at CBS is given for 2005.

4.1.2 Projects with funding from other external resources

4.1.2.1 CMOL

The Center for Computational Modelling of Language (Peter Juel Henriksen, Dan Hardt, Matthias Trautner Kromann, Peter Rossen Skadhauge) has received a grant of 4.5 mill. DDK for 3 years from the Danish Research Council of the Humanities. The purpose of the centre is to provide a framework for basic research projects with the aim of developing models for language processing, particularly from psycholinguistic and corpus-based criteria. The centre's research is organised around three projects: "Grammar Acquisition", with the aim of developing methods for automatic language acquisition from large corpora; "Discontinuous Parsing", with the aim of developing methods for representation and processing within the theoretical framework of "Discontinuous Grammar"; and "Interpretation", with the purpose of developing semantic methods for representation and processing of linguistic phenomena both within and across sentences.

CMOL is cooperating with the Indian Institute of Technology (IIT, Kanpur) and Benares Hindu University (Varanasi) about the development of speech technology in an Asian context. By the end of 2004 CMOL received another grant from Danish Research Council of the Humanities for establishing formal network activities.

The project carried out at the centre are funded from various sources and are described under the separate headings: SweDanes, Nortalk, Representation of transcribed speech, Danish Grammar Checking Systems, The Danish Dependency Treebank (DDT), Dependency Annotation tool (DTAG).

4.1.2.2 Talking Head Project

I arranged a visit of the developers of the RUTH animated dialog system, Matthew Stone and Doug DeCarlo, in May 2004. Together with Stone and DeCarlo, PJH and I have implemented a Danish-speaking version of the system. This system has the potential to form the basis for a variety of research and application-oriented projects involving animation and dialog.

4.1.2.3 SDMT

The project Statistical Dependency-Based Machine Translation (SDMT) is a joint project by the Department and CMOL and the Center for Language Technology at the University of Copenhagen. SDMT aims at the development of basic tools and methodologies for statistical machine translation based on a parallel Danish-English dependency treebank.

The project started in January 2004 and is financed by The Danish Research Council of the Humanities with a grant of 1.8 mio DKK for a period of three years. The project is managed by Sabine Kirchmeier-

Andersen. Jakob Elming is associated as a ph.d. student from 2005 investigating methods for word and sentence-alignment and automatic derivation of transfer rules.

The cost of the original project description is estimated to 8 mio DKK. The grant of 1.8 mio DKK by the Danish Research Council of the Humanities allowed for a reduced project consisting of a ph.d.-grant and the development of the basic resource: the parallel dependency treebank. Although both institutions are contributing a considerable amount of research time, more resources will be needed to complete the original project. During 2004 several attempts were made to raise additional funding.

The treebank is based on the Danish annotated Parole text which has been further annotated with dependency information by the Danish Dependency Treebank (DDT) project at CBS/CMOL. The Danish Parole text is translated into English and annotated automatically with dependency information using an English LFG-parser.

4.1.2.4 OntoQuery

The purpose of the interdisciplinary research project Ontology-based Querying (<http://www.ontoquery.dk>) is the development of theories and methods for content-based information search in textual databases. Within the project, work is performed in parallel on the subjects of search, ontology, syntax/semantics, and prototype development.

The OntoQuery project is supported by the Danish Science Council, within the information technology programme, for the period 1999-2004, with a supplementary grant for part of the financing of a PhD studentship in effect from Maj 1st 2002. The project periods has been extended to July 2005.

The project has participation from: The Intelligent Systems Laboratory at Roskilde University (4 people), Informatics and Mathematical Modelling at the Technical University of Denmark (3), Centre for Language Technology (2), Business Communication and Information Science, and the Department of Computational Linguistics, CBS (5). The participants from the department are Per Anker Jensen, Bodil Nistrup Madsen, Ekaterina Mhaana, Lone Bo Sisseck (maternity leave), and Hanne Erdman Thomsen who is active in the management of the project.

During 2004, particular work has been done on further development of the ontological grammar. Descriptions with the ontological grammar form the basis for evaluation of queries; in particular, measures for closeness or "similarity" between descriptions, and "calculations" with these goals, have been in focus also in 2004

At the Department of Computational Linguistics in particular, work has been done on automatisation of the ontology construction, including formalisation of the inheritance of characteristic features and automatic extraction of relations, and the question of which semantic relations are relevant in an ontology that is to be applied as described in the project. Work concerning the integration of the generative ontological framework with the static ontology in the NL lexicon resource SIMPLE, are continued.

The project hosted an international PhD course on Concept Analysis and Concept Based Retrieval.

During 2004, the project has presented papers at international conferences, and continues to participate in the European OntoWeb network (<http://ontoweb.aifb.uni-karlsruhe.de/>).

4.1.2.5 STO (*SprogTeknologisk Ordbase*)

STO (*SprogTeknologisk Ordbase*) is a project with the aim of developing a 50.000 word computational lexicon for Danish, including morphological, syntactic and to a certain extent semantic information, for use in the development of language technology applications. The project was initiated by the Danish Ministry of IT- and Research's working party "IT in Danish", who had set aside DDK 8 mill. for the project

for 3 years. The work was coordinated by the Centre for Language Technology, and was completed by the end of February 2004.

In 2004 a major effort has been undertaken to convert the STO electronic dictionary into a computational lexicon based on LFG in the Pargram project (cf. section 4.1.2.1).

4.1.2.6 SweDanes

This project deals with the comparative linguistics of Danish and Swedish speech, and is carried out in collaboration with the Dept. of Linguistics, University of Göteborg, Prof. Jens Allwood, Prof. Elisabeth Ahlsén, among others. The project is funded by NorFA. Peter Juel Henriksen participates from the department. (See also NordTalk below.) The project was finished in 2004.

In the course of the project 6 articles were accepted or published in Scandinavian and international journals. Furthermore, a number of working papers appeared such as Copenhagen Working Papers in LSP 3/2003. The project is currently working on a book on Danish and Swedish spoken language with a clear data-driven point of view summing up the achievements of the past 3 years. Articles have been published in International Journal of Corpus Linguistics and Nordic Journal of Linguistics.

4.1.2.7 NordTalk

This project deals with the establishment, exchange and utilization of speech corpora. Representatives of all the Nordic countries, as well as Estonia, participate. Peter Juel Henriksen participates from the department.

Within the two NorFA projects NordTalk and SweDanes, Peter Juel Henrich has developed an algorithm for automatic (n -gram based) translation of spontaneous speech (in orthographic transcription). The method, which aims at application in speech technology, is corpus-based and uses as its basis two transcription corpora (both >250.000 words) in two different, but related languages. The output is a 1:1 translation of the most high frequency words (up to rank 300, which will include the function words in particular). Experiments have been performed with the large Swedish corpus Göteborg Corpus of Spontaneous Speech and the large Danish corpus BySoc. Preliminary results are promising: Among the 100 most frequent Swedish speech words, approximately 90% are translated correctly (for instance, *måste* -> *skal*) and the rest near-correctly.

Peter Juel Henriksen has also worked on methods for automatic grammar annotation; methods for semantic representation based on dynamic logic; statistical/corpus-based comparison between written and spoken Danish, and Danish-Swedish comparative studies in speech/writing; and algorithmic methods for automatic translation between Danish and Swedish spontaneous speech (the formalisms Siblings and Cousins).

In relation to the projects NordTalk and SweDanes, a collaboration with The Danish Centre for Dyslexia has been established, regarding the establishment of a sound based dictionary directed at the needs of dyslectic people.

The results of the project have been published in Copenhagen Working Papers in LSP.

4.1.2.8 FOVITS

Research based continuing education in language and information technology. (Forskningsbaseret efteruddannelse i sprog og informationsteknologi). Funded by the ministry of science, technology and development and carried out in cooperation with CST and Aalborg University. Members of the board Per Anker Jensen and Stig W. Jørgensen.

4.1.3 Research projects in collaboration with The Danish Centre for Terminology

The overall objective of the Danish Centre for Terminology (DANTERMcentret) is to contribute to the development of Danish know-how within the fields of terminology and language technology, and to develop methods and tools for the creation and management of corporate term bases. Through the Centre, the department has improved its potential for establishing good and constructive collaborations with a number of Danish companies. The department has made an active contribution to the efforts to secure the continued existence of the centre, including participating in meetings and seminars and filing applications for funding of joint research projects.

Bodil Nistrup Madsen has, as part of her work for the centre, worked on the principles of content, structure and functionality of a web-based corporate database, i-Term. Work has been carried out on methods and tools for the development of an internet-based lexical database and a website for Blinkenberg & Høybye's Danish-French and French-Danish dictionaries (the project has received DKK 50.000 from Hedorfs Fond).

4.1.3.1 The IT Terminology Project

The purpose of this project is to establish a database and a web page with advice on Danish IT terminology. The project is carried out in collaboration with the Danish Centre for Terminology, the Danish Language Council, the Institute of Computer Science at the University of Copenhagen, the IT University of Copenhagen, the Danish IT Industry Association, a.o. The project has been funded by VILLUM KANN RASMUSSEN FONDEN (DKK 200.000 in 2004).

During 2004, Bodil Nistrup Madsen has participated in work on the development of a new web-based database, and has contributed to principles of content, structure and functionality (principles of feature specifications and concept systems). The database has been programmed by Progresso, and has been published on the Danish Language Council's web page (<http://www.it-dansk.dk>) during spring 2004.

4.1.3.2 Database for the Lykeion Thesaurus

This project deals with information about central concepts in relation to system analysis and construction. In 2003 the following activities have been carried out: analysis, proposal for database structure and conversion of data. In 2004 draft proposals for user interfaces have been worked out. Bodil Nistrup Madsen is in charge of this project.

4.1.4 Projects with funding from internal resources

4.1.4.1 Centre for Terminological Ontologies (CTO)

The work of the Centre for Terminological Ontologies includes basic research as well as application oriented research and development. It is the purpose of CTO to work with methods, which can form the scientific basis for the development of ontology-related IT tools. Research topics include:

- formal terminological concept analysis
- concept analysis based on descriptions of concepts in natural language
- principles for construction of ontologies, with a view to automatic construction of ontologies
- Till now six projects have been planned. However, some projects have not yet been initiated.

Ongoing projects are described below.

4.1.4.1.1 CAOS

CAOS, or Computer-Aided Ontology Structuring (Bodil Nistrup Madsen, Hanne Erdman Thomsen and Carl Vikner, Bo Kranz Simonsen, Jakob M. Christensen), aims at the development of a system for semiautomatic construction of concept systems by means of feature structures, on the basis of user-typed information. In 2003, the project has received DDK 126.640 from the department's research allocation and DDK 90.000 from the foundation *Hedorfs Fond*. The funding has been spent on

programming assistance, release and salary for emeritus professor Carl Vikner who is very active in the project.

In 2004, work has carried out on methods for handling the insertion, deletion and movement of concepts and partial hierarchies. The CAOS system has been demonstrated on several occasions at workshops and seminars. Danish as well as international publications have been prepared.

The basic research concerning principles, necessary for the development of a first prototype of the CAOS system, has been finished in 2003. However, the implementation of some of these principles in the system still remains. Till now only generic concept relations have been covered, but the project 'Formal terminological concept analysis', cf. below, will include research on methods for the handling of other concept relations.

During 2004 the development of graphic facilities based on UML (Unified Modeling Language) has begun.

4.1.4.1.2 Formal terminological concept analysis

The work with CAOS has motivated the formulation of a new project: Formal terminological concept analysis. The purpose of this project is the development of theories and methods for formal description of terminological concept systems (ontologies). In terminological concept analysis it is insufficient to structure concept systems solely by the generic concept relation between superordinate and subordinate concepts. It is also necessary to use such concept relations as for instance the part-whole relation, the causal relation and the resultative relation. Therefore, is it necessary for the formal apparatus to be able to describe such polyrelational concept systems. The theoretical results obtained in the project will be tested and integrated in the CAOS system in the course of the development.

4.1.4.1.3 Web-ontologies

The aim of the project is to investigate and evaluate formalisms and tools designed to build web ontologies, i.e. ontologies, which are meant to support the use of mark up languages on the Internet.

The use of mark up languages, especially the increasing use of XML as a means of mapping content and structure onto tags, does indeed in most cases provide human beings with a nice looking and well structured text as well as a set of tags, the specific vocabulary used to describe the content and structure of the text, easy to understand and, in most cases, easy to use. But not all human beings associate the intended and right meaning with a given tag, and a computer does not associate any meaning at all with a tag. This is one good reason to build an ontology, a mechanism defining classes and relationships among classes and associating classes with properties in a specific domain of interest as an attempt to convert data and tags into meaningful and, hopefully, unambiguous information.

Conceptually, we conceive of a web ontology as a description of concepts and entities and the relationships holding among them as represented in a text or specific area of interest accessible on the Internet and, computationally, as a means of turning data into information and knowledge that can be shared among humans and reused across applications. The aim of the present project is to try to find out how such a device, formally and practically, conceptually and computationally, is defined, designed and used.

Per Anker Jensen joined the group in 2004.

4.1.4.2 Pargram

The goal of the project is to analyse and describe a number of central linguistic phenomena in Danish and Norwegian within the linguistic theory of Lexical-Functional Grammar (LFG) and to implement the descriptions in a state-of-the-art grammar development environment. The goal is, by these means, to produce formal linguistic specifications that will expose differences and similarities between the two languages, and to develop computational linguistic resources for Danish and Norwegian that can be used in actual computational applications, as well as for research and teaching purposes.

The Norfa-funded project ended officially at the end of 2003 but work on the Danish grammar is continued at ID. In 2004 a major effort has been undertaken to convert the STO electronic dictionary into a computational lexicon based on LFG. This work has been carried out in close collaboration with Beau Sheil, who was a visiting researcher at the institute in april and may 2004. At present there is a running version where all the morphological information from STO and a subpart of the STO syntactic descriptions are interfaced with the Danish LFG-grammar. Extensive testing is being carried out. This work has been presented in talks and it is documented in a forthcoming article.

Furthermore an account of VP-topicalization has been developed and implemented in the Danish grammar. The account is based on a novel approach to the analysis of complex verb tenses in LFG and it addresses specific problems relating to various kinds of VP-topicalization in Danish.

4.1.4.3 Danish Grammar Checking Systems

The purpose of the project Danish Grammar Checking Systems (Daniel Hardt, Steffen Leo Hansen, Peter Juel Henriksen) is to develop reliable grammar checkers for a broad range of grammatical problems. A technique has been developed to deduce grammatical principles automatically by means of advanced machine-learning techniques which are applied to syntactically annotated corpora. This technique uses the principles of word class tagging on grammatical problems. Through using a well-known technique in a new way, it becomes possible to develop grammar checkers quickly and with great precision. During 2001 to 2003 the project was funded by the Research Committee.

A website has been developed for the project, and tools have been developed for the automatic generation of grammar checkers. Furthermore, talks have been given and papers have been published on Danish Grammar Checking with Transformation-Based Learning.

The project continues within CMOL and has currently been extended with a writing assistant facility.

4.1.4.4 VIA and Computer Assisted Language Learning

In 1998, the first prototype of the VIA programme was launched with funding from CTU (*Center for teknologistøttet uddannelse*), comprising a total of 1000 exercises for 7 languages. The programme is used in linguistics courses in the IT and English programme, and at high-school level.

In 2003, CBS Learning Lab granted DDK 265.000 for the adaptation of the programme to the web and the development of new exercise types and exercises for the Italian language. The project is led by Bente Lihn Jensen (FIRST, CBS) in cooperation with Sabine Kirchmeier-Andersen, Bo Krantz Simonsen, and Jacob Møller Christensen and CBS Learning Lab.

The reprogramming of VIA for the web including a new design of the interface has been completed, and a general flexible format for the creation of new exercise types has been developed. The programme is now capable of performing automatic morphological analysis and generation by means of a lexicon containing complete morpho-syntactic information for 2500 Italian verbs, nouns and adjectives, which is used for random generation of various exercises and automatic generation of intelligent feed-back. Finally, the project has developed an online grammar for Italian.

In 2004 the official VIA website was established and the programme was made accessible for the public. Discussion. Currently, 13 institutions/departments are using VIA, among these are: CBS, University of Copenhagen, language center of the ministry of foreign affairs and Aurehøj Gymnasium.

The development of VIA will continue as part of the CALL activities at CBS. More info at www.via.id.cbs.dk

4.1.4.5 Multimedia in the Home

The goal of this project is to investigate the possibilities of automatically annotating information about music and making these as well as the corresponding recordings available for everybody who wants to listen to, download, acquire or just be informed about existing recordings of classical music. In the funding period, the project participants have been Mette Nelson, University of Southern Denmark, Kolding, and Steffen Leo Hansen of the department, who is in charge of the project

The work on the prototype FYNBO has been resumed and reorganised in autumn 2004 by Steffen Leo Hansen. In this second phase of the project the focus is on knowledge and meaning as the basis for a description of core of the system. A main feature is the development of a domain specific ontology, the design of a database containing world knowledge about artists and ensembles performing classical music, and the optimisation of a semantic parse of the users' question as a basis for the generation of an answer.

The project will aim at establishing contacts to the university of Münster and the project MUSIL, Münster Semantic Interoperability Lab.

4.1.4.6 The Department's Corpus Initiative

To ensure an optimal utilisation of the department's IBM-sponsored servers, and to ensure the application of results from the ParaT project and others projects that involve corpora, the department's staff work on establishing joint corpus resources at the Faculty of Languages, Communication and Cultural Studies. This work has been formalised in the Department's Corpus Initiative (Steffen Leo Hansen, Peter Juel Henriksen, Dan Hardt, Anders Thøgersen, Henrik Selsøe Sørensen, Sabine Kirchmeier-Andersen, Tina Nielsen).

The initiative concerns organising and making available corpus material on the servers, installing XKWIC and BNC's SARAH as search applications, and morphosyntactically tagging the Danish texts. In the period covered by the report, the first web interface that gives access to all the department's corpora, has been completed. The application was made available to the Faculty of Languages for research and teaching in the spring of 2002.

The number of corpora has been substantially increased, and intensive work has been done on the establishment of a Danish tree bank (see DDT below).

The department has been among the originators of the idea of establishing a Danish Language Bank (*Dansk Sprogbank*) for securing a standardised availability of Danish corpus resources and other resources for developing language technology. No government funding has been made available for this project.

4.1.4.7 ParaT

The project Parallel Texts (Sabine Kirchmeier-Andersen) covers the development of methods for automatically parallelising texts, and tools for the handling of parallel texts, establishment of parallel corpora and research into the use of parallel texts in computational linguistics, for instance in translation memory systems. Work on the further development of the corpus has been at a standstill while the effort has been put into the development of a web interface for displaying parallel corpora on the internet, and a robot for the automatic collection of corpora.

The work has been continued in the DDT project and the SDMT project.

4.1.4.8 Discourse Treebank

Also begun the Fall semester is the **Discourse Group**, as a result of a decision in the ID institutseminar, that discourse should receive a high priority. A group (consisting of Dan Hardt, Peter Juel Henriksen, MTK, Peter Rossen Skadhauge, Per Anker Jensen, and Bjarne Ørsnes) has begun meetings. While we are still in a preliminary planning and study stage: one major goal is to produce a Discourse Treebank for Danish, building upon previous projects such as the RST Discourse Treebank and the ongoing Penn Discourse Treebank. The meetings are open and the group was recently joined by Hans Dybkjær (Prolog Development Center).

4.1.5 Other research projects

4.1.5.1 Stanlex

Bodil Nistrup Madsen heads a project concerning classification and structuring of lexical data. The project was started in 1995 and has the purpose of developing a taxonomy and a description of the content of lexical data in databases and natural language systems, as well as models for the structural description of lexical data using SGML. Hanne Erdmann Thomsen joined the group in 2003.

The work is performed within the framework of a working party under the Danish Standards Association, consisting of people working with research and development within lexicography, terminology, computational linguistics and other fields.

4.1.5.2 Representation of transcribed speech

In collaboration with the NordTalk project, the Dept. of Linguistic, University of Göteborg, and the Department of Danish Dialectology, University of Copenhagen, Peter Juel Henriksen is working on a tool for translation between the most widespread Scandinavian formats for representation and transcription of spontaneous speech – currently in particular the formats CorDiale, MSO06, BySoc and Danish Standard 2.

A grammatically annotated (PAROLE tagged) version of BySoc (large Danish speech corpus) has been worked out, and parsing and web publication has been applied to the large Danish dialectal speech corpus CorDiale (well over 1 mill. running words of spontaneous speech), and corpus TT2 (Danish sentences in verified phonetic transcription for use in speech technology).

In 2003 the project DanPO was initiated with the purpose of providing a transcription of the words in the STO database in order to enable STO to serve as a reference resource for Danish speech technology. The first application will be a talking head for Danish which is being developed at CMOL. Version 1.0 was released with 87.000 transcribed lemmata which are currently being tested in an application of Danish speech synthesis. Papers published at the 2nd Conference on Indian Lexicography and NODALIDA-2005.

Sound transcription and syntactic tagging of the corpus "Spontan-tale" (Spontaneous speech). The project, which is carried out in collaboration with the Department of Applied Linguistics at the University of Copenhagen (Nina Grønnum), comprises 20 speakers and approx. 25000 running words. The tagging is carried out at CMOL.

4.1.5.3 Dependency annotation tool (DTAG)

Matthias T. Kromann is developing a dependency-based syntactic formalism with algorithms for parsing and learning a lexicon from a treebank. The theory is tested in practice in a programme (DTAG) to be used for building large treebanks. The programme is expected to handle: (a) manual dependency annotation of large corpora; (b) display of dependency graphs as arc graphs; (c) constraint-based search for syntactic structures in a dependency treebank; (d) automatic dependency annotation by means of local optimality parsing, from a given lexicon; (e) automatic construction of a lexicon from a word class hierarchy and a dependency treebank annotated with types from the word class hierarchy; and (f) "error highlighting" that indicates where a dependency graph contradicts an underlying lexicon, for use in manual verification of a treebank.

The project is a natural continuation of M.T. Kroman's former PhD project on Discountinuous Grammar. 6000 lines of programming code has been written, making DTAG a fully usable tool with respect to manual dependency annotation (a), display of dependency graphs (b), and syntactic search (c). This part of DTAG has formed the basis for the tagging of the first 60.000 words in DDT. Furthermore, algorithms have been developed for automatic parsing and lexicon building. Results can be seen at <http://disgram.sf.net> and <http://www.id.cbs.dk/~mkt/dq>.

The DTAG algorithms for searching, parsing, and grammar learning formed the starting point of the Statistical Dependency-based Machine Translation project (SDMT).

The project is part of Matthias post-doctoral dissertation which is expected to be finished in 2005.

4.1.6 PhD Projects

The department has five PhD students.

Tina Nielsen investigates computer assisted learning applications for teaching language technology. The purpose of the investigation is to formulate, discuss and test construction principles for computer assisted learning in relation to the teaching of LSP and language technology at university level. Tina Nielsen's PhD project is financed by the Danish Research Council for the Humanities and affiliated to the Danish Centre for Terminology. Tina has left the department after the end of her grant period without finishing her ph.d.

Nina Sværke Hansen (formerly Frederiksen) works on a contrastive study of the phenomena of "extraction" in Danish and French, with the aim of implementation in Lexical-Functional Grammar (LFG). Nina Sværke Hansen is affiliated with the PARGRAM project. Nina has been on maternity leave until august 2004.

Ekaterina Mhaanna works on the formal properties of ontologies. She is affiliated with the OntoQuery project. Ekaterina has been on maternity leave in 2004.

Lone Bo Sisseck investigates methods for identification and extraction of conceptual relations in Danish domain-specific texts in order to be able to generate ontologies automatically. It is a very time consuming task to build an ontology manually. It will therefore be of great interest to be able to construct ontologies automatically or semi automatically. Earlier and ongoing projects throughout the world have already developed various methods of more or less automatic term and relation extraction from both a statistical and a linguistic approach. However, there have hardly been any attempts to investigate Danish texts in order to find linguistic patterns that could indicate relations between concepts and eventually serve as a tool to build ontologies automatically.

Her project is primary financed by the Danish Technical Research Council and is affiliated to the OntoQuery project. Her project is also supported with additional means from the Copenhagen Business School.

Jakob Halskov was given a PhD grant by the Faculty in 2003 to work on the detection of new words and terms in parallel corpora. Probing the Properties of Determinologization is a corpus-based study of the semantic shifts which occur when non-specialists apply terminology from the domain of Information Technology in general language discourse. The project implements a so-called DiaSketch which can be used to detect the degree of determinologization of a given term in a given corpus. While the DiaSketch is inspired by the lexical profiling techniques developed by Adam Kilgarrieff, it is new in that it introduces a longitudinal and terminological perspective. Overall, the project aims at increasing our understanding of how clear-cut concepts may deteriorate into fuzzy categories over time.

Jakob Elming is associated as a ph.d. student from 2005 investigating methods for word and sentence-alignment and automatic derivation of transfer rules.

4.2 Other academic activities in 2004

The department's staff are academically active in a large number of ways and have participated in many seminars, conferences and courses. Information about these activities is only included to the extent that they fall within the following five headings: PhD study programme, external activities, guest lectures, conferences and course development. Activities such as participation in CBS committees, course responsibility and coordination, supervision and evaluation of job applicants are not included.

4.2.1 PhD study programme

The department has participated in a joint application for the establishment of a Nordic Graduate School of Language Technology. An application was submitted to NORFA in June 2003. Funding has been granted in 2004. Dan Hardt is a member of the board, representing Denmark. <http://ngslt.org/>

The department is actively contributing to the national graduate programme, GradEast.

Sabine Kirchmeier-Andersen is the main supervisor for Matthias Trautner Kromann and for Jakob Halskov.

Hanne Erdman Thomsen is main supervisor for Lone Bo Sisseck.

Bjarne Ørsnes is supervisor for Nina Sværke Hansen.

Bodil Nistrup Madsen is supervisor for Ekaterina Mhaanna.

Per Anker Jensen is co-supervisor for Tine Lassen

The department's staff have

- arranged PhD courses under the aegis of GradEast, including syntax and morphology
- arranged and taught a PhD course on representation formalisms for ontologies at an international graduate school (FQAS)
- taught at the European Summer School in Logic, Language and Information (ESLLI)
- arranged PhD courses under the aegis of Nordic graduate School in Language Technology

4.2.2 Network activities

The department is part of the computational linguistics network in Denmark, consisting of all computational linguistics environments. The purpose of the network is an exchange of research results and coordination of the computational linguistics research.

The department is represented in six international research networks:

- Nordic Network for Lexical Grammar (NorFA)
- The European OntoWeb network, Ontology-based information exchange for knowledge management and electronic commerce (EU)
- Nordic treebank network (Norfa)
- Nordic CALL network – computer assisted language learning (Norfa)
- ScanMatrix – network for grammar development and machine translation (NorFa).

The department has entered negotiations for moving the international Association for Terminology and Knowledge Transfer, GTW (*Gesellschaft für Terminologie und Wissenstransfer*), to Copenhagen (Dept. of Computational Linguistics/The Danish Centre for Terminology).

4.2.3 International conferences

The department has hosted international and Scandinavian conferences and workshops in connection to the described projects.

Furthermore, GTW's next international conference TKE 2005 (Terminology and Knowledge Engineering) will be held in Copenhagen. See <http://gtw-org.uibk.ac.at/>

4.2.4 Participation in interdisciplinary work groups and committees

- Membership of the Danish Standards Association's work group on health informatics, SUSI AG 2
- Chairmanship of the Danish Standards Association's STANLEX working group
- Chairmanship of ISO TC37 SC3
- Chairmanship of Danish Terminology Group

- Membership of the coordination group of the IT Terminology Committee
- Membership of expert team in CEN (European Committee for Standardization) concerning standards for e-Cataloguing (Multilingual catalogue strategies for ecommerce and ebusiness)
- Participation as consultant in the terminology work of the National Board of Health and *Hovedstadens Sygehusfællesskab* (Copenhagen hospitals)
- Participation in a dialogue between the ministries of science and culture, research institutions, and a large number of companies on the subject of establishing language technology resources for research and the business community, including the establishment of a "Danish Language Bank"
- Membership of the committee of the Lykeion society, an interest organisation which, among other things, has the aim of developing models and other tools for the benefit of professional communication across professional fields. The members are, among others, several large firms of consulting engineers, the Danish Agency for Trade and Industry, staff from the National Board of Health, the Royal School of Library and Information Science, Denmark, the Technical University of Denmark and other research institutions
- Membership of the committee, and hosting of the annual meeting of LEDA (association of Danish lexicographers)
- Membership of the CBS Center for Continuing Education and Business Research

4.2.5 Participation in collaborative projects

- Collaboration with Software AG about the XML database Tamino (sponsored in 2001). Preliminary discussions about joint research and education projects
- Collaboration with The Danish Centre for Dyslexia on the establishment of a sound based dictionary directed at the needs of dyslectic people
- Collaboration with the Danish Center for Terminology on business oriented projects
- Collaboration with Mikroværkstedet A/S on student exchange and development of CALL-programmes.

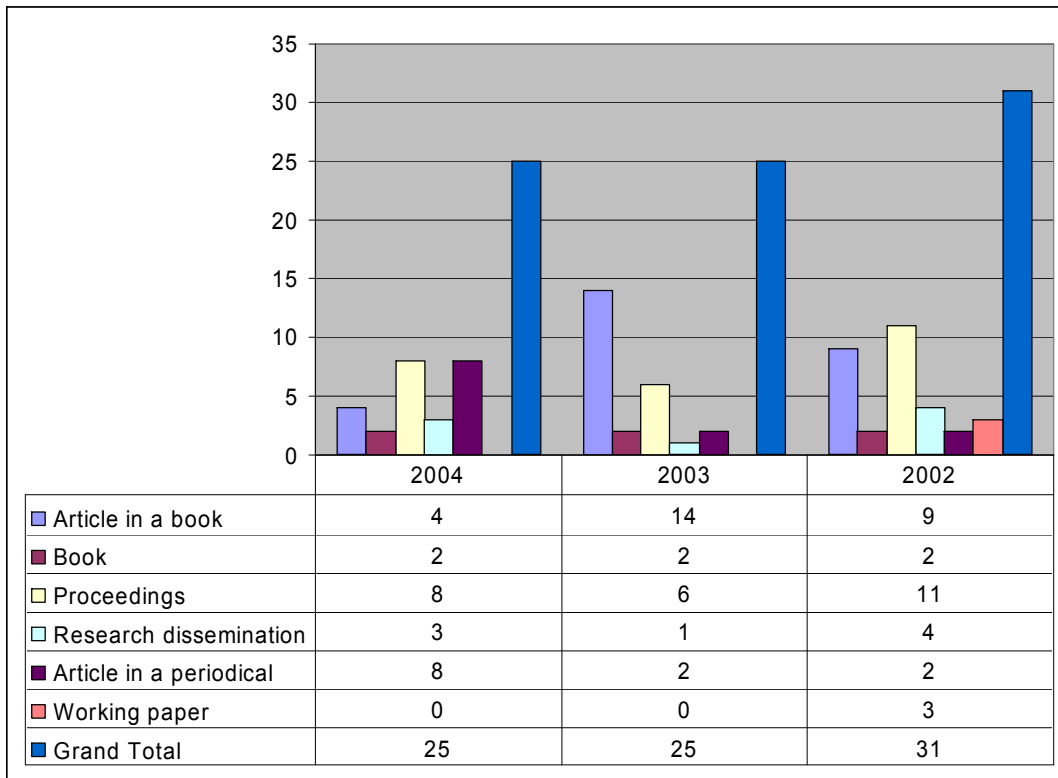
4.2.6 Evaluation and reviews

The department's staff have performed various evaluation tasks, including

- Evaluating research projects for The Norwegian Research Council
- Reviewing for Computational Linguistics and Association for Computational Linguistics European Chapter
- Membership of the scientific committee for the international LREC conference in 2004. Reviewing conference papers
- Membership of the editorial board of Nordic Journal of Linguistics
- Reviewing for ACTA Linguistica Hafniensis

4.2.7 Publications

The table below compares the number of publications 2002 - 2004. The number of publications in 2004 is the same as in 2003. However, in 2004 there are more publications in English and more reviewed articles in periodicals, and less articles in books than in 2003. (A few articles did not make it in time for the official statistics, these have been included into the statistics of the present report).



5 Teaching activities

Teaching activities are described in detail by the programme directors of the various programmes in separate reports. This section gives an overview of the teaching activities of the department. It should be stressed that the department strives that all teaching activities are research based, since we consider teaching the best way to disseminate research results. By research based we thus mean courses which in addition to leading theories in the field also present recent research results of the department, either by the researchers himself or by a qualified teacher instructed by the researchers.

5.1 The Master's programme in Computational Linguistics

In 2003 the department has published a qualifications profile for the master's programme in computational linguistics. The profile documents the intended occupation of the candidates, the qualifications required to carry out this occupation and how the programme aims at providing the candidates with these qualifications. In addition the department has continued its work on implementing the new Master's programme in Computational Linguistics. The programme will be revised after the BA specialisation in language technology has been launched

5.2 MA supplementary course in Language Policy

The department has developed a course for students in the master's programme in modern language and business communication. The content of the specialisation is the design and implementation of language policy in companies with a special emphasis on the IT tools available to support this work. The course pays special attention to localisation and was offered for the first time in autumn 2003. In 2004 9 students attended the course.

5.3 The BA programme in Language Technology and English

A lot of work has been put into giving the programme its final form. The basis of this work was the experience gained from the programme IT and English, and a questionnaire made by the evaluation unit. The results can be viewed at <http://www.cbs.dk/studies/sprogtek/undervisning>.

The revised programme started in September 2003. A key issue of the revision was to equip students with good skills within text production and communication in Danish and English, and within the use and evaluation of language technology. At the same time, efforts were made to increase and maintain the students interest in the programme. This was first and foremost done by reducing the number of obligatory disciplines and increasing the number of short IT-related course. Secondly, more focus was put on areas of language technology such as mark-up languages, speech technology, machine translation and natural language processing.

During 2001-2004 the number of students enrolled in the programme has dropped with more than 50%. This is most likely due to the fact that interest in IT has dropped world wide and consequently, that students within the Faculty of Language, Communication and Cultural Studies are shifting from more technical areas to more general humanist areas such as communication and intercultural studies.

This development is rather alarming considering the fact that language technology will continue to be important and that the BA feeds not only the MA programme in computational linguistics but also other programmes at for instance the IT University.

The department has a three-line strategy to overcome this situation:

- extending the programme to cover all languages taught at the faculty.
- developing a new programme, BA in knowledge management and communication, combining the areas of communication, English and computational linguistics (see below).
- putting more effort into the marketing of language technology and computational linguistics in close cooperation with the communications unit of CBS.

In September 2005 Language Technology will be integrated in the BA in international business communication.

5.4 BA in Management of Information, Communication and Knowledge (MICK)

In cooperation with the Center for Communication, the English Department, the Department of Informatics and CBS Learning Lab a new programme is being developed comprising the following areas:

- communication: general communication, corporate communication etc.
- language technology: content management, document management, knowledge organisation, classification, markup etc.
- English: high degree of professional bilingualism, special regard for parallel information and knowledge processing in Danish and English

5.5 The Master of Language Administration programme

In 2003 the MLA-programme was moved to the SITESCAPE learning space and has now only 4 seminar days per Semester. All teaching activities are carried out in the learning space in small self-organising working groups. The MLA programme has been associated with CBS CELL (Center for executive learning and leadership).

In 2004 the MLA programme became a part of the Erasmus Mundus programme LATER. During 2005 it will be transformed into an international distant learning programme and offered in English.

5.6 Initiatives in further education

Carsten Hansen has developed and held courses on language technology for high school teachers in autumn 2004. This was made possible by a grant from the faculty. The department is interested in further development in this area, however, currently there are no resources for these activities.

The course material is available at <http://www.id.cbs.dk/~ch/gymkurs/>

This report was written and edited by Sabine Kirchmeier-Andersen and Stig W. Jørgensen