1. Project Title

Preparing DutchSemCor: A Dutch corpus with word senses from the Cornetto database

2. Summary

The *startsubsidie* will be used to prepare an NWO *middel-groot investeringsproject* to be submitted in September 2008. The goal of that final project is to deliver a corpus that is fully sense-tagged with senses, ontology tags, and domain tags from the Cornetto database. This corpus will play a key role in language technology research for Dutch and also in linguistic and cognitive research that relates linguistic form to meaning. Combining the best of both worlds, the corpus will be tagged using a combination of automatic techniques and manual editing. Automatic tagging techniques include on the one hand supervised methods, which can be trained on already tagged subcorpora as training data, enabling them to tag other subcorpora, and on the other hand unsupervised techniques that rely on other sources such as the Cornetto database itself. It is to be expected that the manual editing of the corpus will feed back in the form of adaptations to the semantic database Cornetto.

3. Main applicant

Prof. Piek Vossen, LCC, Faculteit der Letteren, Vrije Universiteit Amsterdam, De Boelelaan 1105, Tel. +31 (0)20 5986466, Fax. +31 (0)20 5986500, email: p.vossen@let.vu.nl

4. Composition of the research group and contact person

Applicant	Short form	Address	Contact	Email	Tel.
Language, Cognition and Communication, Faculteit der Letteren, Vrije Universiteit Amsterdam	VUA	De Boelelaan 1105, 1081HV, Amsterdam	Prof. Piek Vossen	p.vossen@let.vu.nl	020 - 5986466
Communicatie- en Informatie-wetenschappen, Faculteit Geesteswetenschappen, Universiteit van Tilburg	UvT	Postbus 90153, 5000 LE Tilburg	Dr. Antal van den Bosch	Antal.vdnBosch@uvt.nl	013 - 4663117
Instituut voor Informatica, Universiteit van Amsterdam	UvA	Kruislaan 403, 1098 SJ Amsterdam	Prof. Maarten de Rijke	mdr@science.uva.nl	020 – 5255358

The consortium represents a combination of 3 different disciplines:

- Automatic word sense disambiguation (UvT, UvA)
- Corpus annotation technology (UvT)
- Linguistic-lexicographic expertise (VUA)

5. Institutional setting

The work of the main applicant VUA will be carried out within the focus group *Language, Cognition and Communication* (LCC) of the Faculty of Arts. The professor chair Computational Lexicology is a member of LCC. The research of LCC is part of the interfaculty research institute *The Center for Advanced Media Research Amsterdam* (Camera) at the VUA, in which the Faculty of Arts, the Faculty of Exact Sciences, the Faculty of Social Science and the Faculty of Psychology and Pedagogy take part.

6. Period of Funding

8 months

7. Description of the proposed project

7.1 Groups and partners involved

The consortium represents top researchers in the areas of word sense disambiguation technology and automatic corpus tagging (UvT and UvA), and in the area of computational lexicology (VUA). Furthermore, the consortium is also active in developing higher-level natural language processing and information processing applications such as question-answering systems, dialogue systems, and search engines, that can exploit such a corpus directly.

7.2 Added value of the proposed cooperation

The goal of the final proposal will be to deliver a corpus that is fully sense-tagged with senses and domain tags from the Cornetto database.[1] This much-desired corpus will play a key role in language technology research for Dutch (until now, only one small and domain-specific sense-annotated corpus for Dutch exists in the research domain), and also in psycholinguistic, sociolinguistic, and cognitive research in which the relations between linguistic forms and their meaning are investigated. The corpus will be tagged using a combination of automatic techniques and manual computer-assisted editing. Automatic tagging techniques include on the one hand supervised methods, which can be trained on already tagged subcorpora as training data, enabling them to tag other subcorpora automatically, and on the other hand unsupervised techniques that rely on other sources such as the Cornetto database itself. It is to be expected that the manual editing of the corpus (both from scratch and correcting the output of the automatic techniques) will feed back in the form of adaptations to the semantic database Cornetto.

7.3 Workplan for the *startsubsidie*

We will investigate the following issues:

1. Selection of the corpus

[1] Cornetto is a project funded by the STEVIN programme of the Dutch Language Union with NWO, FWO, IWT, and SenterNovem; see http://www.let.vu.nl/onderzoek/projectsites/cornetto/introduction.html

2. Different types and levels of semantic tagging
3. Encoding protocols and standards for *manual* semantic tagging
4. Quality of *automatic* tagging systems
5. Speed and efficiency of *automatic versus manual* semantic tagging
6. Possible uses and exploitation of a Dutch Semcor and the requirements of size, quality and type of tagging

The preparatory work funded by the *startsubsidie* would be divided over the following tasks:

1. (Month 1-3) Study of state of the art with respect to:
    a. semantically tagged corpora,
    b. semantic tagging systems and methods,
    c. encoding schemes and standards for semantic tagging,
    d. Dutch corpora and annotations, and
    e. language technology exploiting semantically tagged corpora.
2. (Month 5) Organisation of a workshop on semantic tagging where international experts are invited, a.o. Dr. Fellbaum (Princeton University), Dr. A Kilgarriff (Lexicography MasterClass Ltd.), Dr. E. Agirre (Basque Country University), Dr. McCarthy (University of Sussex).
3. (Month 6) Writing the specifications for a semantically tagged corpus of Dutch
4. (Month 7) Pilot to manually assign semantic tags to a small corpus according to the specifications.
5. (Month 8) Writing of the proposal.

As part of the work we expect to have 4 meetings within the consortium: 2 meetings before the workshop and 2 meetings after the workshop. We will also set up Twiki and Trac websites for the project where project members can collect information, data, and specifications, share ideas, and manage their joint activities.

7.4 Coordination of research agendas

Following the workplan outlined above, the teams will collaborate on writing a state-of-the-art report and use these efforts both to share their research agendas and to coordinate them. Further coordination will be realized while detailing the specifications and running the planned pilot.

7.5 Added value for the research community

SemCor[2] and other semantically tagged corpora have been available for English for many years now. These corpora have been instrumental in building supervised word sense disambiguation (WSD) systems, which have shown their applicability in higher-level NLP and information processing systems (e.g. question-answering, text generation, information extraction, and semantic role labeling).

A Dutch sense-tagged corpus would not only be essential for the development of general-purpose Dutch WSD systems; it would also leverage the Dutch text mining toolkit (including

[2] SemCor, containing 700 thousand sense-tagged words from the Brown corpus, was created at Princeton University with senses from WordNet developed by the same group; current versions of SemCor can be downloaded freely from http://www.cs.unt.edu/~rada/downloads.html#semcor

also parsers and named-entity recognition systems developed by other groups) and automatic concept and ontology acquisition tools. On top of that, since the concepts are also related to English concepts in the Princeton WordNet, machine translation methods could be developed that use this corpus for aligning meanings and linguistics expressions with sense-tagged corpora in other languages.

From a more linguistic-theoretical perspective, a sense-tagged corpus could be used to provide empirical input to long-standing issues on the relation between linguistic form and meaning, especially with complex phenomena such as metaphors, and meaning extension, generalization and specialization.

7.6 Planned grant application

The *startsubsidie* will be used to prepare an NWO *middel-groot investeringsproject* to be submitted in September 2008.

8. Curriculum vitae of the main applicant

Prof. Vossen studied General Linguistics at the University of Amsterdam, with minors Language Philosophy and Cognitive Science. After his graduation in 1986, he worked for 12 years at the University of Amsterdam in the area of Computational Lexicology, mainly in national (Links, Like) and European projects (Acquilex, Links, EuroWordNet). In 1995, he received his PhD (Cum Laude) in Computational Lexicology at the University of Amsterdam. He was the initiator and coordinator of the EuroWordNet project. At the end of that project he founded the Global Wordnet Association, together with Dr. Christiane Fellbaum from Princeton University. From 2000 onwards he has been working for the language technology companies Sail Labs (Antwerp) and Irion Technology (Delft). He currently is the CTO of Irion Technologies. Since 2007, he is professor of Computational Lexicology at the Faculty of Arts at the Vrije Universiteit Amsterdam, where is he is the coordinator of the Stevin project Cornetto. The Cornetto project will result in a major Dutch semantic database that combines the Dutch WordNet with the Referentie Bestand Nederlands. In 2008, he will be coordinating the 7th Framework project KYOTO that aims at a global semantic encoding and knowledge representation system related to different languages and cultures.

9. Project Budget

	Euro	Motivation
Project leader	**15.000**	**Studying state of the art, coordinating pilot annotation project, writing the proposal**
Workshop	**10.000**	**Inviting international experts**
Student assistants	**5.000**	**Performing annotations in pilot annotation project**
Technical experts	**5.000**	**Writing the guidelines and specifications**
Total	**35.000**	

Prof. Vossen will be the project leader. He is assigned at the VUA exclusively for research projects, and he is available within the current research program of the Faculty.

10. References

Agirre, E., and Edmonds, P., editors (2006). *Word Sense Disambiguation Algorithms and Applications*. Text. Speech and Language Technology Series, Vol. 33. Dordrecht: Springer. ISBN: 978-1-4020-4808-1

Fellbaum, C., Palmer, M., Hoa Trang Dang, Delfs, L., and Wolf, S. (2001). Manual and Automatic Semantic Annotation with WordNet. In *WordNet and Other Lexical Resources, NAACL 2001 Workshop*, Pittsburg, PA.

Hoste,V., Hendrickx, I., Daelemans, W., and Van den Bosch, A. (2002). Parameter optimization for machine learning of word sense disambiguation. *Natural Language Engineering*, **8:4**, pp. 311-325

Jijkoun, V. and M. de Rijke. Recognizing Textual Entailment: Is Lexical Similarity Enough?, In: I. Dagan, F. Dalche, J. Quinonero Candela, B. Magnini, editors, *Evaluating Predictive Uncertainty, Textual Entailment and Object Recognition Systems*, LNAI 3944, pp. 449-460, Springer Verlag, May 2006.

Killgariff, A. (1999) 95% Replicability for manual word sense tagging**.** In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, Bergen, Norway, pp. 277-278. ACL.

Vossen, P., Hofman, K., De Rijke, M., Tjong Kim Sang, E., and Deschacht, K. (2007) The Cornetto Database: Architecture and User-Scenarios, In: *Proceedings of 7th Dutch-Belgian Information Retrieval Workshop DIR2007*, Leuven: University of Leuven.

Vossen P. (2002) *EuroWordNet General Document.* EuroWordNet Project Report LE2-4003 & LE4-8328. Amsterdam: University of Amsterdam.