DECLARATION:

I, Gideon Jozua Kotzé, hereby declare that this study is my own work and that it has not previously been submitted for assessment to another University or for another qualification.

SIGNATURE: **……..**

DATE: **……..**

Building a WordNet for Afrikaans:

Preliminary research in the form of an inquiry into the feasibility and optimal methodology for the development of a wordnet database

Gideon Kotzé
Sherada 13
Van Riebeeckstraat
Potchefstroom
2531
Zuid-Afrika
tel: +27 834480651
Studentnummer: 1548565

Vrije Universiteit
Faculteit der Letteren
Opleiding Lexicologie en Terminologie

Begeleider: Prof. Dr. P. Vossen
Tweede lezer: Dr. H. van der Vliet

December 2006

Abstract

This thesis is written as part of the preliminary research for a proposed project at the Centre for Text Technology at the North-West University in Potchefstroom, North-West Province, South Africa. In this work a methodology for constructing a wordnet for Afrikaans is proposed, which will be based on the Princeton WordNet that was developed at Princeton University, USA. All relevant concepts are introduced, starting with an analysis of the prototype wordnet, the Princeton WordNet, after which the focus shifts to its extension to multilingual wordnet databases. An investigation is made into the available resources and tools available for the proposed project, after which we verify its feasibility. Afterwards, various methodologies for wordnet construction are investigated and analysed. Based on all these analyses, a detailed methodology with a work schedule is proposed for building a core Afrikaans wordnet, also keeping in mind future extension and potential problems. We conclude that the existence of several automatic techniques have greatly improved the process of wordnet construction compared to a few years ago, but that they are still heavily dependent on quality lexical resources and tools. Finally, some suggestions are made for future extensions and applications of the wordnet.

KEYWORDS: Afrikaans wordnet project, South Africa, Princeton WordNet, lexical database, semantic network, wordnet construction methodology, natural language processing, synset, semantic relation, lexical relation, equivalence relation, expand methodology.

ACKNOWLEDGEMENTS

I would gladly like to acknowledge and thank the following persons who have aided me in one way or another in the production of this thesis:

- Prof Piek Vossen, my supervisor, and wordnet expert supreme, who has provided me time and again with invaluable help and who has agreed to act as international collaborator for the Afrikaans wordnet project.
- Dr Hennie van der Vliet, my second reader and lecturer in various courses during my M.A. studies at the Free University in Amsterdam, for his patience, great sense of humour and continuous encouragement. His great interest in, among other areas, lexical semantics and its computational applications was contagious and has definitely been a factor in choosing this topic.
- Prof Dr Willy Martin, recently retired Head of the Department of Lexicology and Terminology of the Free University of Amsterdam and lecturer in various courses during my studies there as well, for his friendliness, personally caring about his students and his ability to effectively transfer information from his great pool of knowledge in a way that challenges and inspires; and for the chance he gave me to work on Cornetto, introducing me to Prof Vossen and the fascinating world of lexical databases.
- Prof Gerhard van Huyssteen from the Centre for Text Technology at the North West University in Potchefstroom, South Africa, for believing in me enough to employ me at the Centre for, among other things, work on the Afrikaans wordnet.
- Prof Karel Pala and Dr Ales Horak from Masaryk University in Brno, Czech Republic, for providing me with valuable information regarding the wordnet development tools VisDic and DEBVisDic.
- My family, for their great support and care throughout and the opportunities that they gave me.
- My friends, for having to hear time and again that the thesis is not finished yet (!), and of course, their support as well.
- My Creator, for giving me the talent, intelligence, motivation and strength to engage and complete this work; also for surrounding me with loved ones who supported me all the way.

TABLE OF CONTENTS

LIST OF FIGURES

Figure number and name **Page number**

INTRODUCTION

1. Background

In the realm of lexicography, computational lexicology and also in other branches of linguistics such as psycholinguistics, the structure, nature and accurate representation of human language and knowledge, as well as the relationship between the latter two, has long been a topic of extensive research and remains a great challenge today. The advent of computer technology has made it possible to process continually greater quantities of data, also those describing the abovementioned knowledge, and in ever more innovative ways. As most aspects of language can be represented as a complicated structured whole, or at least as parts that have some relation to the whole, it is only logical that language can quite adequately be represented as data structures by and on computer systems. It follows that certain kinds of processing can also be applied to them. It is these processing capabilities of computer systems that have led to much fruitful research helping to understand the nature of human language and also dealing efficiently with applications such as translation, lexicography and intelligent processing techniques. One result of this is the lexical database, of which the wordnet is a prime example.

Working on a Dutch lexical database, Cornetto, during my studies in Lexicology and Terminology in Amsterdam, The Netherlands, has piqued my interest in this particular discipline. Because of my being appointed a position at the North West University in Potchefstroom, South Africa, as researcher and data developer working on, among other things, a wordnet for Afrikaans, it seemed appropriate to conduct the required preliminary research in the form of this thesis.

The aim of this thesis, then, is to provide an answer as to how to build a wordnet for Afrikaans most efficiently, considering the current literature, used methodologies and practical considerations such as available resources.

2. Thesis methodology

Besides providing background information about WordNet, as well as describing the specific project environment, the rest of the thesis consists of finding and describing an optimal methodology that is to be used to build a wordnet for Afrikaans. This information will also mainly come from literature, the details of which follow under the literature review section, but might also include other sources such as personal contact with experts and own insight resulting from the unique situation surrounding this particular project. Data to be studied and/or used for the purpose of this work and the project include lexicographic sources and software tools. The nature of the research is mainly qualitative, but conclusions and decisions may also rely on calculations and statistics where the analysis of different methodologies is concerned.

Firstly, a thorough background will be given in chapter 1. This concerns the background of the project, as well as the most important resource of all, the Princeton WordNet. We shall discuss how it came into being, what it is, what is its purpose, its evolution, advantages and limitations, after which multilingual wordnet databases such as EuroWordNet and BalkaNet are presented. Then we discuss an organisation that plays a very central role today in the wordnet scene, the Global WordNet Association.

In chapter 2, we shed some light onto the context in which our project is taking place, looking briefly at all relevant factors: the staff, the amount of resources, including lexicographic, computational and financial, and how we generally might proceed. The facts stated here are used to make a conclusion on the feasibility of the proposed project.

Before any suggestions can be made, we first need to look at a few different methodologies that were applied in the building of other wordnets. In chapter 3 we follow this process, comparing them to each other and commenting on the procedures involved.

In chapter 4, a methodology is selected that is based on the facts presented in chapters 2 and 3. We describe the implementation of the chosen methodology for this particular project and also include a work schedule. The factors to be considered here

are the lexicographical resources, the software tools, the hardware, the staff, the time frame and possible future development for the wordnet (for example, expansion to other languages or integration with other systems). If the methodology is imperfect for our particular situation, suggestions will be made for improvement.

Expected results and problems will be presented in chapter 5. Then a summary will be given of the research results with a relevant discussion. Finally, conclusions are drawn and suggestions are made for future work in this field, mentioning some details concerning our intentions and aims using the future Afrikaans wordnet as a resource and an application.

3. Literature review

WordNet is, among other things, a practical result of many decades of accumulated linguistic research. It is inspired by theories of relational lexical semantics, where, basically, meaning is depicted by networks of nodes consisting out of words or word groups that are connected to each other through specific lexical or semantic relationships. This stands in contrast with the older school of componential lexical semantics, where words are seen to be decomposable into smaller semantic abstract units, similar to the way a sentence can be broken down. George Miller discusses these two schools in relation to the WordNet structure, in the foreword of the book *WordNet: An Electronic Lexical Database* (Fellbaum 1998:xvi). Some important works on relational lexical semantics and also lexical semantics in general are:

- Cruse, D.A. 1986. *Lexical Semantics*. Cambridge, England: Cambridge University Press.
- Jackendoff, R. 1983. *Semantics and cognition*. Cambridge, MA: MIT Press.
- Pustejovsky, J. 1995. *The generative lexicon*. Cambridge, MA: MIT Press.
- Levin, B. (Ed.). 1985. *Lexical semantics in review*. Cambridge, MA: MIT, Center for Cognitive Science.
- Evens, M. (Ed.) 1988. *Relational models of the lexicon.* Cambridge, England: Cambridge University Press.
- Mel'čuk. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In: Leo Wanner, 1996. *Lexical Functions in Lexicography and Natural Language Processing*, Benjamins, Amsterdam.
- Sowa, John. Lexical Structures and Conceptual Structures. In: *Semantics and the Lexicon*, 223-262. 1993. Kluwer Academic Publishers. Printed in the Netherlands.

A great book for those relatively uninitiated in the fields of semantics and, by implication, lexical semantics, is *Semantics* by John I. Saeed[1] (2003). Even though it assumes some linguistic knowledge, it is clearly written and treats the most important concepts and current theories relevant to the field.

[1] Saeed, John. 2003. *Semantics*. Malden, MA: Blackwell Pub.

The paper *Semantic Networks* by John F. Sowa[2] (http://www.jfsowa.com/pubs/semnet.htm) defines and explains different kinds of semantic networks. One of them, among others like assertional and learning networks, is the *definitional network*, of which WordNet is an example. More detail about this follows in chapter 1.

Probably the most important work regarding WordNet is the book *WordNet: An Electronic Lexical Database*, which was mentioned earlier. Called "a landmark book" by Dagobert Soergel of D-Lib Magazine[3], this should be compulsory material for anyone who wants to get acquainted with this phenomenon. Soergel defines the target audience as "anyone interested in language, in dictionaries and thesauri, or in natural language processing". Written in great detail, it not only describes how it came into being, what it is and how it works, but also devotes a number of chapters on various kinds of applications that WordNet can be used for. This book has become a top reference work on WordNet and has, not surprisingly, played a very important role in the research for this work.

The predecessor of this book was an influential series of five papers published in the *International Journal of Lexicography* in 1990[4]. Although the first few chapters in the abovementioned WordNet book are meant to replace these papers, which are now considered partially outdated, Soergel states in the above review that the five papers "…give more detail and interesting discussions not found in the book."

---

[2] The first paragraph states that it is a revised and extended version of an article that was originally written for the *Encyclopedia of Artificial Intelligence*, edited by Stuart C. Shapiro, Wiley, 1987, second edition, 1992.

[3] Soergel, Dagobert. WordNet. In: *D-Lib Magazine*. October 1998. http://www.dlib.org/dlib/october98/10bookreview.html)

[4]

- Miller, George et al. *Introduction to WordNet: An On-line Lexical Database*.
- Miller, George. *Nouns in WordNet: A Lexical Inheritance System*.
- Gross, Derek. and Miller, Katherine. *Adjectives in WordNet*.
- Fellbaum, Christiane. *English Verbs as a Semantic Net*.
- Beckwith, Richard and Miller, George. 1990. Implementing a Lexical Network. In: *International Journal of Lexicography* 3(4), 1990, pp. 302-312.
- All of the above in: *International Journal of Lexicography* 3(4), 1990.

The November 1995 issue of *Communications of the ACM* published an article[5] by George Miller, which we find to be a good summary, what the PWN is and does. It also mentions some problems, such as contextual representations and polysemy, and suggestions for dealing with them.

A very insightful discourse can be found in the paper *WordNet and EDR: Critiques and Responses*[6] between George Miller, Doug Lenat and Toshio Yokoi regarding three different systems: WordNet, CYC and EDR respectively. On a web page of the *Communications of the ACM*[7], the abstract of an article[8] is shown, describing CYC as follows: "Cyc is a bold attempt to assemble a massive knowledge base (on the order of 108 axioms) spanning human consensus knowledge." The official web site describes it as follows:

> The Cyc knowledge base (KB) is a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life.

As can be derived from these statements, CYC is a knowledge base using artificial intelligence to infer truths and non-truths from a series of statements or axioms. WordNet, on the other hand, is rather an attempt at an electronic representation of the mental lexicon and its semantic and lexical relationships. EDR is defined as follows on its official web site[9]:

> The *EDR Electronic Dictionary* was developed for advanced processing of natural language by computers, and is composed of eleven sub-dictionaries. Sub-dictionaries include a concept dictionary, word dictionaries, bilingual dictionaries, etc (…) The *EDR Electronic Dictionary* is a machine-tractable dictionary that

---

[5] Miller, George A. WordNet: A Lexical Database for English. In: *Communications of the ACM*, Volume 38, Issue 11 (November 1995). Pp. 39-41

[6] Lenat, D., Miller, G. and Yokoi, T. CYC, WordNet and EDR: Critiques and Responses. In: *Communications of the ACM*, Volume 38, Issue 11 (November 1995). pp. 45-48.

[7] http://portal.acm.org/citation.cfm?doid=79173.79176

[8] Lenat, Doug et al. 1990. Cyc: toward programs with common sense. In: *Communications of the ACM,* Volume 33, Issue 8 (August 1990). It is downloadable on the same page, at http://portal.acm.org/ft_gateway.cfm?id=79176&type=pdf&coll=GUIDE&dl=GUIDE&CFID=5537083&CFTOKEN=95991261.

[9] http://www2.nict.go.jp/r/r312/EDR/index.html

> catalogues the lexical knowledge of Japanese and English (…) and has unified thesaurus-like concept classifications (…) with corpus databases.

In chapter 2 there is a section on why we chose to build a wordnet instead of another system like CYC.

Jorge Morato and Miguel Angel Marzal, at the Global WordNet Conference of 2004, present a clear and insightful paper on how WordNet has been utilised thus far[10]. All the important general applications are listed and discussed. Some useful papers on the various applications of WordNet are:

- Katz et al. *Word Sense Disambiguation For Information Retrieval*. http://people.csail.mit.edu/ozlem/abstract00-uzuner-wsd.pdf
- Rosso et al. *Text Categorization and Information Retrieval Using WordNet Senses*. http://www.fi.muni.cz/gwc2004/proc/110.pdf
- Buscaldi et al. 2005. *A WordNet-based Query Expansion method for Geographical Information Retrieval*. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/buscaldi05.pdf
- Mandala et al. Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. In: *Proceedings of EACL '99*. http://delivery.acm.org/10.1145/980000/977049/p94-mandala.pdf
- Rigau G., Agirre E. Disambiguating bilingual nominal entries against WordNet. In: *Workshop On The Computational Lexicon - ESSLLI, 71-82. Barcelona. 1995*. http://arxiv.org/PS_cache/cmp-lg/pdf/9510/9510004.pdf
- Kwong, O. Aligning WordNet with Additional Lexical Resources. Coling-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems, 1998, Montréal, Canada. http://www.ai.sri.com/~harabagi/coling-acl98/acl_work/olivia.ps.gz.

The Princeton WordNet was only the beginning of a global phenomenon. Other language communities have noted its success and started to build wordnets of their

[10] Morato, J., Marzal, M., Lloréns, J. and Moreiro, J. WordNet Applications. In: *GWC 2004, Proceedings*, pp. 270-278. Sojka, Petr, Pala, Karel, Smrž, Pavel, Fellbaum, Christiane, Vossen, Piek (eds.). ©Masaryk University, Brno, 2003. www.fi.muni.cz/gwc2004/proc/105.pdf.

own. Of course, the benefits of having a multilingual database comprising several different wordnets for the purpose of cross-linguistic research and application are obvious. EuroWordNet (EWN) is one example of such a database. As some techniques and approaches that were used in this project are important for our endeavour, the relevant literature is also given ample attention in this work. The following are some important papers on various aspects of the structure of the EWN:

- Vossen P. (1999) *EuroWordNet General Document*. EuroWordNet LE2-4003, LE4-8328. 1999. http://www.hum.uva.nl/~ewn/docs/GeneralDocPS.zip
- Nancy Ide, Daniel Greenstein, Piek Vossen (eds.). *Special Issue on EuroWordNet. Computers and the Humanities*, Volume 32, Nos. 2-3 1998. 117-152.
- Vossen, P. (ed.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.
- Climent, S., Rodríguez, H. and Gonzalo, J. 1996. *Definition of the links and subsets for nouns of the EuroWordNet project.* Version 6. EuroWordNet LE2-4003. http://cv.uoc.es/~grc0_001091_web/files/climent96.pdf.

Various other documents on EWN can be found on the EuroWordNet website, http://www.illc.uva.nl/EuroWordNet/docs/.

Over time, the idea of EWN has evolved into an endeavour to create a multilingual wordnet database for all languages in the world. This ideal is being promoted by the Global Wordnet Association (GWA). The idea consists of using the so-called Inter-Lingual-Index (ILI), which was introduced by EWN, to connect all existing wordnets, making some semi-automatic translation possible and greatly facilitating cross-linguistic reference and research. The ILI is, to quote from *EuroWordNet: General Documentation*:

> (…) an unstructured list of meanings (…) where each ILI-record consists of a synset[11], an English gloss specifying the meaning and a reference to its source.

[11] A set of synonyms meant to represent a concept. This is a term peculiar to WordNet and will be explained in greater detail in chapter 1.

> The only purpose of the ILI is to mediate between the synsets of the language-specific wordnets. (Vossen 2002:9)

If the Afrikaans wordnet complies with all required standards, it will in the future be able to link to this ILI, and hence connect to all other languages connected to it. Because the GWA promotes the development of wordnets all over the world, as well as the sharing of information and expertise, being a part of that will prove to bring many benefits and help ensure the success of the Afrikaans wordnet.

Much information about the organisation and their ideals and activities is available on the website of the GWA, http://www.globalwordnet.org. It follows the EWN as a standard and provides data and techniques for the development of wordnets. These are some important documents on the topic:

- Vossen P. (ed.) 2002. *EuroWordNet General Document*. http://www.hum.uva.nl/~ewn.
- Vossen, Piek et al. 1998. *EuroWordNet Tools and Resources Report*. Deliverable D021D025, WP2, EuroWordNet, LE2-4003. http://www.illc.uva.nl/EuroWordNet/docs/D021D025PS.zip
- *Base Concepts*. http://www.globalwordnet.org/gwa/gwa_baseconcepts.htm.
- *EuroWordNet top-ontology*. http://www.globalwordnet.org/gwa/ewn_to_bc/topont.htm.
- *Basic Concepts in Wordnets*. http://www.globalwordnet.org/gwa/ewn_to_bc/corebcs.html.

These papers describe some important data, such as the base concepts, used in EuroWordNet and recommended for use in the building of new wordnets that use the expand methodology. In the second paper (1998), information is provided about the tools and resources used in the project.

There are a lot of papers available that have been published in journals of the proceedings of conferences, many of them freely downloadable on the Internet, about various development methods and tools for building wordnets. Over the years,

methodologies have been refined and other software exists to assist in these processes. Here are just a few of these papers - excluding the abovementioned EWN documents - that are relevant to this work:

- Horák, A., Pala, K., Rambousek, A. and Povolný, M. DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In: *Proceedings of the Second International WordNet Conference – GWC 2004,* pages 136-141, Brno, Czech Republic, 2003. http://nlp.fi.muni.cz/publications/gwc2006_hales_pala_etal/gwc2006_hales_pala_etal.pdf
- *Automatic Building of Wordnets*. Barbu, Eduard and Mititelu, Verginica Barbu. Automatic Building of Wordnets. In: *Proceedings: International Conference – Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgaria. 21-23 September 2005. (https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2005/RANLP/papers/89_barbu.pdf)
- *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets*. Atserias, J., S. Climent, X. Farreres, G. Rigau, and H. Rodriguez, 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In: *Proceedings of "Recent Advances on Natural Language Processing" RANLP'97*, Tzigov Chark, Bulgaria. http://cv.uoc.es/~grc0_001091_web/files/atserias.pdf
- Farreres, X., Rigau, G. and Rodríguez, H. *Using WordNet for Building WordNets*. Universitat Politecnica de Catalunya, Spain. http://www.ai.sri.com/~harabagi/coling-acl98/acl_work/rigau.ps.gz

Following are papers on methodologies used to build specific wordnets:

- Erjavec, T. and Fišer, D. 2006. Building Slovene WordNet. In: *Proceedings of the 5th International Conference on Language Resources and Evaluations, LREC 2006*. http://nl.ijs.si/slownet/bib/slown-LREC05.pdf
- Marrafa, Palmira. 2002. *Portuguese WordNet general architecture and internal semantic relations*. São Paulo.

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502002000300008&lng=es&nrm=iso&tlng=en

- Bentez, L., Cervell, S., Escudero, G., López, M., Rigau, G. and Taulé, M. 1998. Methods and Tools for Building the Catalan WordNet. In: *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources & Evaluation*. Granada, Spain. http://www.lsi.upc.es/~escudero/papers/lrec98.

And finally, these are papers on other multilingual wordnets besides EWN:

- Sinha, M., Reddy, M. and Bhattacharyya, P. *An Approach towards Construction and Application of Multilingual Indo-WordNet*. www.cse.iitb.ac.in/~pb/papers/gwc06_IITB_IndoWN.pdf
- Pianta, E., Bentivogli, L. and Girardi, C. MultiWordNet: Developing an aligned multilingual database. In: *Proceedings of the 1st International WordNet Conference, January 21-25, 2002, Mysore, India, pp. 293-302*. http://multiwordnet.itc.it/paper/MWN-India-published.pdf.
- Tufiş, D., Cristea, D. and Stamou, S. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In: *Romanian Journal of Information Science and Technology*. Volume 7, Numbers 1-2, 2004, 9-43. www.ceid.upatras.gr/Balkanet/journal/7_Overview.pdf.

CHAPTER 1: Defining the concepts: The Princeton WordNet and its offspring

1.1 The Princeton WordNet

1.1.1 *Background, theory and structure of the Princeton WordNet*

The Princeton WordNet (PWN) was developed at the beginning of the 1990's by the Cognitive Science Laboratory at Princeton University, USA, under the direction of Professor George A. Miller (Principal Investigator). It is a free lexical reference database, based on a psycholinguistic model of the mental lexicon. The basic lexical data are sets of synonyms, or as it is more commonly known, synsets, consisting of content words in their base forms: nouns, verbs, adjectives and adverbs. Together with semantic relations between these synsets, they form the basic structure of WordNet. The database has proved extremely useful for its accessibility, quick reference and potential for serving as a base or support for other language technological and lexicographical applications. The original idea was to organise knowledge in a similar way to how it is thought to be organised in the human brain (Fellbaum 1998:2). The developers used results from psycholinguistic research to help determine aspects of the structure of their project. For example, their use of lexical relations instead of conceptual relations to link adjective antonym pairs are based on empirical research, the reason being that speakers tend to associate these pairs based on their word forms rather than conceptually (Fellbaum 1998:9). Later, the database became more important for computational linguistics, but of course also for lexicography and theoretical research in especially semantics. As stated in the introduction, wordnets in many other languages have been developed since then, eventually resulting in fully-fledged multilingual wordnet databases such as EuroWordNet.

The PWN is a lexical database in the relational semantic sense. A lexical database is, according to www.dictionary.com (citing WordNet 2.0) "a database of information about words". It stores, structures and presents knowledge, in the form of natural language on word level, in a systematic way. As stated in the introduction, relational semantics focus on the representation of meaning in a network-like structure, where

the nodes are semantic units, possibly word forms, connected to each other via semantic or lexical relations. The wordnet is a very specific form of this kind of lexical database, of which the PWN is the grand prototype. A formal definition of a "wordnet" is provided by www.dictionary.com:

> n 1: any of the machine-readable lexical databases modelled after the Princeton WordNet 2: a machine-readable lexical database organized by meanings; developed at Princeton University [syn: WordNet, Princeton WordNet]

A lexical database can have a multitude of uses, such as being a repository for lexical information in order to produce lexicographic works like dictionaries, providing valuable statistical data about a language (usually with the aid of corpora and additional software), or being actively used as a tool in natural language processing (NLP). The latter includes playing a role in important functions such as machine translation, where techniques like word sense disambiguation are very important, or information retrieval on the Internet or other data sources.

The on-line article, *Semantic dictionary viewed as a lexical database* by Paducheva et al (1992), defines the purpose of a lexical database in the following way:

> The purpose of a [lexical] database is to highlight those semantic aspects of a word that unite semantically cognate words and differentiate many of semantically different words from one another. In other words, [a] lexical database is an instrument of predicting and calculating all sorts of useful semantic classes of words.

The focus in WordNet is on lemmatised or base word forms and concepts, which are represented as sets of synonyms, briefly called synsets, as mentioned earlier.

It is already stated in the introduction that WordNet is a type of *definitional network*. John F. Sowa explains the term in his paper, *Semantic Networks* (1992)[12]:

[12] The first paragraph states: "This is a revised and extended version of an article that was originally written for the *Encyclopedia of Artificial Intelligence*, edited by Stuart C. Shapiro, Wiley, 1987, second edition, 1992."

> *Definitional networks* emphasize the *subtype* or *is-a* relation between a concept type and a newly defined subtype. The resulting network, also called a *generalization* or *subsumption* hierarchy, supports the rule of *inheritance* for copying properties defined for a supertype to all of its subtypes. Since definitions are true by definition, the information in these networks is often assumed to be necessarily true.

The structure of WordNet, at least with regard to some relation types, is therefore mainly hierarchical. For example, *car* is a kind of *vehicle*, and thus the concept *car* is represented in a node underneath the one representing the concept *vehicle*, inheriting all the properties of *vehicle* and containing more semantic information, i.e. it is more specific. This kind of relation is called *hyponymy* and is one of the basic relations in WordNet.

As mentioned earlier, the concepts represented in WordNet are categorized into nouns, verbs, adjectives and adverbs. A few possible reasons that function words are not included are that they are composed of relatively small closed word classes and that they probably cannot be represented adequately through a wordnet-like structure. Some of these classes would also be difficult to organize in any kind of semantic network: A word like *the*, for example, has an almost purely grammatical sense. From a multilingual point of view, they can also tend to be very language-specific, where a function word in one language can, for example, be represented in the other by one or more morphemes. The content words are grouped according to their word classes. For the most part, relations do not cross word class boundaries: Nouns refer to other nouns, verbs to other verbs, etc. Of course, this also applies to noun phrases, verb phrases, etc. in case they are represented each as one element in a synset. Concepts are represented by way of so-called *synonym sets*, shortened to *synsets*. Here is an example of a synset:

- **enter**, come in, get into, get in, go into, go in, move into

This list of words refers to only one sense of the word *enter*. Here is another one:

- **enter**, participate

It is obvious here that the above two synsets refer to two different senses of *enter*. Every concept is made explicit by formal definitions and, optionally, example sentences or phrases. This is the full entry of the second synset, using the default display of the online browser:

- **enter**, participate (become a participant; be involved in) *"enter a race"; "enter an agreement"; "enter a drug treatment program"; "enter negotiations"*

There are two kinds of relations in WordNet: *lexical* and *conceptual-semantic* (often shortened to *semantic*). Lexical relations link individual words and conceptual-semantic relations link concepts (i.e. synsets) (Fellbaum 1998:9). As mentioned earlier, an example of a lexical relation is the antonym relation. The fact that the words *fast* and *slow* are generally considered antonyms and not, for example, *rapid* and *slow* suggests that this relation, albeit still a semantic one (Fellbaum 1998:49), is word form based.

There are quite a few different kinds of relations. Some of them are specific to one word class, such as *troponyms* to verbs; others are shared by more than one of them, such as *hypernyms* (nouns and verbs) and antonyms (adjectives, adverbs and verbs).

When you enter a word in the WordNet user interface to search the database, a list of all synsets is displayed in which the word occurs. If the word appears in more than one word class, the synsets from those classes are named as such and kept apart. One can traverse the list of all possible relations from a specific word or synset by choosing it and selecting the appropriate option, the details of which depend on the browser used (on- or offline). Navigating through these relations enables one to find and display the concepts or word forms which are linked to the previously displayed concepts or word forms through this relation. In this way, one can quickly get an idea of the general (mostly) paradigmatic structure of the English lexicon and its concepts.

Figures 1 and 2 show tables with all the relations in WordNet (excluding synonymy) with a brief definition and example[13].

Noun	**Definition**	**Verb**	**Definition**
Antonymy	Indicating word forms with opposing values, such as "victory" and "defeat".	Antonymy	Also labelled as "opposition", verb antonyms usually denote different thematic roles of the same action, e.g. "buy" and "sell" (1998:81)
Hyponymy	"[hyponymy is] the relation of subordination (or class inclusion or subsumption)" (1998:24), e.g. "robin" is a hyponym of "bird". The reverse is called the "hypernym"[14], e.g. "bird" is the "hypernym" of "robin".	Troponymy	"The troponymy relation between two verbs can be expressed by the formula *To V_1 is to V_2 in some particular manner*." (1998:79). E.g. "whisper" is a troponym of "talk". The converse is called the "hypernym", e.g. "talk" is a hypernym of "whisper".
Meronymy	"The part-whole relation between nouns is generally considered to be a semantic relation, called *meronymy*" (1998:37). E.g. a meronym of "car" is "wheel". The reverse is called the "holonym", e.g. "car" is a "holonym"	Entailment	"*Entailment* is used here to refer to the relation between two verbs V_1 and V_2 that holds when the sentence *Someone V_1* logically entails the sentence *Someone V_2*. For example, *snore* lexically entails *sleep* because the sentence *He is snoring* entails *He is sleeping*…" (1998:77)

[13] These two tables are a combination of a table of relations found in (Fellbaum 1998:109) and various definitions, most of which are found in the same work. Quotes are put in quotation marks.
[14] WordNet 2.1 gives "hypernym" as the correct spelling. However, several works refer to the concept as "hyperonym". We have chosen the spelling as it is displayed in WordNet, except for where we quote, in which case we do not regard "hyperonym" as a spelling error, because of its wide-spread use.

	of "wheel".		
Attribute	"Values of attributes are expressed by adjectives. For example, SIZE and COLOR are attributes of robins: the size of robins can be described by the adjective *small* (…) the color (…) by (…) *red*." (1998:40) E.g. under "speed", the attribute relation refers to the possible attribute values, "fast" and "slow".	Cause	"The cause relation picks out two verb concepts, one causitive (like *give*), the other what might be called "resultative" (like *have*)." (1998:83)
		Also see	Self-explanatory. For example, it could refer to a derivationally related form, such as the verb "smoke" to the noun "smoke".

Figure 1: Table of WordNet relations and their definitions for nouns and verbs

Adjective	**Definition**	**Adverb**	**Definition**
Antonymy	"Antonymous adjectives express opposing values of an attribute. For example, the antonym of *heavy* is *light*, which expresses a value at the opposite pole of the WEIGHT attribute." (1998:48)	Antonym	"Adverbs derived from adjectives frequently inherit from their related adjectives such properties as antonymy (…) For example, the antonymous relation between the adjectives *specific* and *general* is found also between the related adverbs, *specifically* and *generally*." (1998:60-61)
Similar	Self-explanatory. For example, the adjective "chicken" has a link, "similar to" or "synonyms/related nouns", which links to, among others, "cowardly" and "fearful" which have similar meanings but in	Derived from	"Most adverbs belong to the large open class derived from adjectives by suffixation. Of these, the great majority are derived by adding the suffix *–ly* and typically specify manner

	this case are used in a different register.		(*beautifully, oddly, quickly* (…))" (1998:60). Therefore, "beautifully" links to the adjective "beautiful" through this relation.
Relational adj.	"adjectives that are related semantically and morphologically to nouns (…) For example, *musical* in *musical instrument* is related to *music* and *dental* in *dental hygiene* is related to *tooth*" (1998:59).		
Also see	Self-explanatory. For example, the adjective "fast" links with this relation to the words "expedited", "hurried" and "sudden", which are related but cannot be considered as synonyms.		
Attribute	The adjective "fast", for example, links with this relation to the noun "speed", which is an attribute of which "fast" gives a certain value.		
Participle	"Most of these adjectives are the participle forms of verbs. The adjectives in *an obliging waiter, elapsed time*, and *his accustomed thoroughness* are what we are calling *participial adjectives*." (1998:58) Therefore, "elapsed" links to the verb "elapse" through this relation.		

Figure 2: Table of WordNet relations and their definitions for adjectives and adverbs

Apart from these relations, there are other categories as well. Some entries are classified according to a specific domain in which they occur. For example, under *central processing unit* there is link called *domain category*, which links to the entry *computer science*. Domain categorisation further helps to link words to specific senses

through disambiguation, as well as make studies possible on topics like the vocabulary of subject language.

Sometimes, words are classified according to their specific style or register. For example, under the adjective *chicken*, there is a link called *domain usage*, which refers to the entry *colloquialism*. Another link occurring here is *similar to*, which refers to words (*cowardly, fearful*) that more or less mean the same as the entry word (*chicken*), but which are used in a different style or register.

Many adjectives have the link *attribute* that refers to a category that the adjective characterises. For example, *fast* is linked to the entries *speed, swiftness* and *fastness* through the *attribute* relation. One of the synsets in which the word *fast* appears also contains the link *see also*, which refers to synsets that are closely related but cannot be regarded as synonyms. Examples are *expedited, hurried* and *sudden*.

As mentioned earlier, relations are structured differently in different word categories. For example, nouns are mostly strictly hierarchical, where, on the other hand, adjectives and their antonyms are organised in clusters in a bipolar structure. The reason is that antonymy is the basic relation among descriptive adjectives[15] (Fellbaum 1998:48) but some adjectives do not have antonyms. Such words are then grouped together with a word that do have an antonym and where all of these words form part of one synset. The words that do not have antonyms are then classified as indirect antonyms of the direct antonymic word of the head of the group (Fellbaum 1998:50).

1.1.2 *Advantages of WordNet*

WordNet has many useful features and advantages in the worlds of natural language processing, computational linguistics and lexicography, of which the following are undoubtedly just a few:

- It is one of the world's leading and most widely used electronic lexical-conceptual databases.

[15] WordNet divides adjectives into two categories: descriptive, which is the largest category, and relational, which are related by derivation to nouns (Fellbaum 1998:47).

- The word form and its meaning or meanings, as well as the relationships between word forms, between meanings, and between the former and the latter, are represented in a simple, effective and standard manner, ideal for human and machine.
- It is freely available, meaning that anyone can use the data. One can access it on-line without even downloading it, although that option is also available.
- The WordNet package includes various tools, such as a morphological analyser. The latter is called "Morphy".
- It is easily integrated with other systems. Examples will be given in section 1.1.4.
- In the article, *From WordNet to a Knowledge Base* (2006), Clark et al state that "WordNet is attractive to use because of its comprehensive coverage, syntactic simplicity, comprehensibility, availability, ease of use, and semantic organization".
- There is a huge amount of published literature freely available on the Internet. A lot of research from all over the world has contributed to a rapid improvement and expansion of the PWN. Various other databases has been created using WordNet as a base, such as VerbNet[16].
- WordNet has a multitude of applications. They will also be discussed in section 1.1.4.
- There are, at the time of writing, 47 different wordnets in the world[17]. Many of them are integrated in multilingual wordnet projects, such as EuroWordNet, BalkaNet and Indo WordNet. There is also a great amount of on-line literature available on these projects. This greatly facilitates the studying of different wordnet building methodologies, making it easier to build a wordnet based on good, proven methods and principles.
- The Global WordNet Association, at the time of writing under the chairmanship of Dr Christiane Fellbaum of Princeton University, USA, aims to help maintain, standardise and link all the wordnets in the world to each other[18]. The GWA builds on the results of the PWN and EuroWordNet.

---

[16] See: Palmer, M.: Consistent criteria for sense distinctions. *Computers and the Humanities*, 34 (1-2) (2000) 217-222.
[17] http://www.globalwordnet.org/gwa/wordnet_table.htm
[18] http://www.globalwordnet.org

1.1.3 *Current limitations of WordNet*

The PWN does have its fair share of limitations. Although this and the other problems mentioned here may change after the time of writing, currently there is very little on syntax and nothing on morphology, pronunciation, the forms of irregular verbs or the etimology. The approach is mostly paradigmatic, that is, describing the relationship between the word form and/or the concept that it represents and those that can replace it through semantic or lexical relations. Overall, very little attention is given to collocations and selection restrictions, in other words, how words typically interact with each other, and which combinations are acceptable and which are not. To compensate, definitions and example sentences or phrases are added. In the case of verbs, so-called *sentence frames* are provided where the basic syntactic order is given, where the constituents include *somebody* or *something*, as, for example, under one of the synsets in which *receive* appears:

This is the synset:

> S: (v) **receive**, have (get something; come into possession of) *"receive payment"; "receive a gift"; "receive letters from the front"*

And this is the sentence frame:

- Somebody ----s something
- Somebody ----s somebody
- Somebody ----s something from somebody
- [Applies to receive] The banks receive the check
- [Applies to receive] They receive the money
- [Applies to receive] They receive more bread
- [Applies to have] The banks have the check
- [Applies to have] They have the money
- [Applies to have] They have more bread

The “s” that is added here is the only morphological information provided in WordNet. With adjectives, glosses are used in an attempt to provide typical nouns that

can combine with the adjectives. Unfortunately, these combinations (lexical or semantic) are not displayed in a systematic way.

When one types in a word that deviates from its base form, one is directed to the entry of the base form. This is done with a morphological analyser called Morphy (1998:124) that is part of the WordNet suite of programmes, which in turn is called LexPert (Beckwith and Miller 1990:302). On the user level there is no morphological information available. The focus is purely on the word form and its relationship with the concept and vice versa.

Clark et al (2006) state "A frequent criticism of WordNet is that is has 'too many' word senses." In other words, the sense distinctions are very fine. This could result in it being difficult in some situations to choose the right sense, as well as to link to another wordnet via the Inter-Lingual Index, or to align with another lexical resource, where in both cases the distinctions may be coarser. In the same article, some missing aspects are mentioned that would be desirable in a wordnet used as a knowledge base for machine reasoning. Examples of perceived flaws mentioned here are:

- the dichotomy between verbs and their nominalizations, which are accorded different concepts but, according to the authors, should really be the same. For example, the verb "run" and its noun form ("the run") "denote the same concept (a running event)" (Clark et al 2006)
- a frequently incorrect or dubious hypernym tree
- "unclear or vague" glosses
- missing senses
- the lack of a top-level ontology

Another point of critique is the existence of relatively few semantic relations (see Lenat, Miller and Yokoi 1995:45), which makes the database fall short of processing natural language successfully in comparison with other NLP systems. Contributing to this fact is the lack of background information in order to disambiguate, for example, polysemous words and pronouns (also Lenat, Miller and Yokoi 1995:45-46). Lenat

mentions a few examples such as the following, as a demonstration of this shortcoming of WordNet:

- The pen is in the box.
- The box is in the pen.

Without background information, the right sense of *pen* cannot be determined. The same goes for syntactic ambiguity, with another example from Lenat:

- Fred saw the plane flying over Zurich.
- Fred saw the mountains flying over Zurich.

In this case, a programme needs some background information to be able to determine that the subject of *flying* in the first sentence is *the plane*, while in the second sentence, it is *Fred*, since mountains cannot fly.

In the same article, George Miller admits that "more semantic relations between word senses are desirable" and "much more than a WordNet-type lexical database is required for processing natural language satisfactorily".

However, this article was published a long time ago, in 1995. Since then, WordNet has grown considerably and has also been refined. Today, there are many applications that are integrated with WordNet, exploiting its strong points and compensating for the weak ones, such as the disambiguation problem stated above. WordNet is actually very useful in some disambiguating techniques in conjunction with other sources and/or applications, as stated in the on-line article *Word Sense Disambiguation based on Semantic Density* by Mihalcea and Moldovan[19]. Also, even a part of the CYC database, touted by Lenat in the above 1995 article to be superior in NLP to WordNet, at least with regard to its successful use of background information in word sense disambiguation, is connected to WordNet senses[20], demonstrating the importance of a good semantic base on word level.

[19] www.ai.sri.com/~harabagi/coling-acl98/acl_work/moldovan.ps.gz
[20] More than 6000 links: see http://www.ksl.stanford.edu/onto-std/mailarchive/0171.html.

Harabagiu and Moldovan state in (Fellbaum 1998) that WordNet, in spite of utilising just a few lexical relations, provides a better connectivity than one of the most used machine-readable dictionaries, the *Longman Dictionary of Contemporary English* (LDOCE), but that

> (…) by far, the largest connectivity is found in CYC (Lenat 1995), but at the expense of a weak structure that reduces its applicability to general-purpose reasoning. (Fellbaum 1998:381)

1.1.4 *Applications of WordNet*

Despite its drawbacks, the Princeton WordNet has become one of the most widely used natural language processing resources in existence today. Following is a short summary on its current main applications:

In an article, *WordNet Applications* (2004), by J. Morato, M.A. Marzal, J. Lloréns and J. Moreiro, a figure displays the results of an extensive search of publications on WordNet, which reveals that "the major use of this tool has been in the area of conceptual disambiguation" (Morato et al 2004:271). According to this figure, the other main topics are: improvements in WordNet, image retrieval, machine translation, query expansion, information retrieval and document classification.

With regard to information retrieval, Morato et al mention that "[t]hese operations are closely related to organisation and representation of knowledge on the Internet." (2004:272), stressing the application of artificial intelligence and inferential processes. The authors mention, among others, an article from Moldovan and Mihalcea (2000)[21] who demonstrate how to use WordNet to "optimise the precision of Internet search engines by expanding queries" (2004:272).

Further on, conceptual disambiguation is defined as "precision and relevance in response to a query via resolution of semantic inconsistencies" and is said to be "unquestionably the most abundant and varied WordNet application." (2004:273) In this regard, WordNet has frequently been used in conjunction with other tools or

[21] Moldovan, D.I. and Mihalcea, R.: Using WordNet and lexical operators to improve Internet searchers. In: *IEEE Internet Computing*, 4(1) (2000) 34-43.

systems for improved information retrieval efficiency, as stated in Morato et al (2004:273), for example "for the exploitation of a Bayesian network able to establish lexical relations with WordNet as a source of knowledge, integrating symbolic and statistical information"[22]. The work stated here is Wiebe et al (1998)[23].

Another example mentioned here in the field of disambiguation is "generating ontological databases with a systematic classification of multiple meanings derived from WordNet" (2004:273), referencing a work from P. Buitelaar (1998)[24].

An example of query expansion is given by the authors in mentioning Gonzalo et al (1998)[25], stating that the search process is enhanced "by including semantically related terms and thus retrieve texts in which the query terms do not specifically appear." (2004:274)

Regarding document structuring and categorisation or classification, the authors mention works describing methods to achieve this by "extraction of semantic traits" and "categorisation of the relevance of the data (…) using keywords and WordNet conceptual representation of knowledge". The references given here are Scheler (1996)[26] and Mock, K.J. and Vemuri, V.R. (1997)[27]. Another interesting paper named here is that of Judith Klavans[28] who "devised an algorithm for automatically

---

[22] An on-line tutorial on Bayesian nets and probability, found on a web page of the official site of the Department of Computer Science at Queen Mary, University of London, http://www.dcs.qmw.ac.uk/%7Enorman/BBNs/BBNs.htm, defines a "Bayesian Network" as follows: "A Bayesian Network (also called Bayesian belief network, belief network, Bayesian net, BBN, BN or graphical probability model) is a model for reasoning about uncertainty. Founded on the centuries-old Bayesian probability theory (invented by Thomas Bayes in 1763), the subject has been given a lease of life in recent years due to advances in algorithms and theory. These advances mean that it is now possible to build and run realistic Bayesian nets for a wide range of applications."

[23] Wiebe, Janyce, O'Hara, Tom, and Bruce, Rebecca: Constructing Bayesian Networks from WordNet for Word-Sense Disambiguation: Representational and Processing Issues. *Use of {W}ord{N}et in Natural Language Processing Systems: Proceedings of the Conference Association for Computational Linguistics*, Somerset, New Jersey (1998), 23-30.

[24] Buitelaar, P.: CORELEX: an ontology of systematic polysemous class. *Proceedings FOIS'98*. Amsterdam: IOS Press. (1998) 221-235.

[25] Gonzalo J., Verdejo, F., Chugur, I., and Cigarran, J.: Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP*. Montreal (Canada) (1998) 38-44.

[26] Scheler, G.: Extracting semantic features from unrestricted text. *WCNN'96*. Mahwah (NJ): L. Erlbaum. (1996).

[27] Mock, K.J. and Vemuri, V.R.: Information filtering via hill climbing, WordNet and index patterns. *Information Processing and Management*, 33 (5). (1997) 633-644.

[28] Klavans, Judith and Kan, Min-Yen: Role of verbs in document analysis. *Proceedings of the Conference, COLING-ACL*. Canada: Université de Montreal. (1998).

determining the genre of a paper on the grounds of the WordNet verb categories used" (2004:275).

Other applications mentioned here are those in the domain of audio and video retrieval[29], parameterisable information systems[30] (the example given is "an information system (called Meaning Extraction System) that can be configured in terms of a specific user profile" (2004:275)), language teaching[31] and translation[32].

WordNet is also seen, for example, to be used with corpora for sense identification[33] and aligned with other lexical resources[34].

Of course, there are also several works proposing many kinds of improvements that can be made to WordNet. For example, Clark et al[35] (2006) put forth some suggestions on how to expand WordNet into a fully-fledged knowledge base that can be used more effectively for machine reasoning. Bentivogli and Pianta[36] (2004) show how WordNet can be extended with syntagmatic information. And Wong[37] (2004) points out how some concepts are organized relatively arbitrarily in WordNet 1.5 and

---

[29] Zaiane, O.R., Hagen, E., and Han, J.: Word taxonomy for online visual asset management and mining. Application of Natural Language to Information Systems. Proc. 4th Internat. Conference NLDB'99. Vienna: Osterreichische Comput. Gesseleschaft (1999) 271-275.

[30] Chai, Joyce Y. and Biermann, Alan W.: The use of word sense disambiguation in an information extraction system. Proceedings 16th National Conference on Artificial Intelligence. Menlo Park (Ca): AAAI Press (1999) 850-855.

[31] Hu, X & Graesser, A.: Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in tutorial dialogue. Proceedings of ICCE'98, 2. Beijing: China Higher Education Press (1998) 337-341.

[32] Shei, C.C. and Pain, H.: An ESL writer's collocational aid. Computer Assisted Language Learning, 13 (2) (2000) 167-182.

Diekema, A., Oroumchian, F., Sheridan, P., and Liddy, E.D.: TREC-7 evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. Gaitherburg (USA): TREC-7 National Institute of Standards & Technology (1999) 169-180.

[33] Leacock, C., Miller, G.A., Chodorow, M. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. In: *Computational Linguistics*, 1998, Volume 24, Number 1.

[34] Kwong, O. Aligning WordNet with Additional Lexical Resources. Coling-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems, 1998, Montréal, Canada. http://www.ai.sri.com/~harabagi/coling-acl98/acl_work/olivia.ps.gz.

Mandala et al. 1999. Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. In: *Proceedings of EACL '99*.

[35] P. Clark, P. Harrison, T. Jenkins, J. Thompson, R. Wojcik. From WordNet to a Knowlege Base. In: *Proc AAAI 2006 Spring Symposium on Formalizing and Compiling Background Knowledge*, 2006.

[36] Bentivogli, Luisa and Pianta, Emanuele. Extending WordNet with Syntagmatic Information. In: Sojka, P., Pala, K., Smrž, Fellbaum, C. and Vossen, P. (Eds.) *GWC 2004, Proceedings*, pp. 47-53. © Masaryk University, Brno, 2003.

[37] Wong, Shun Ha Sylvia. 2004. Fighting Arbitrariness in WordNet-like Lexical Databases – A Natural Language Motivated Remedy. In: Sojka, P., Pala, K., Smrž, Fellbaum, C. and Vossen, P. (Eds.) *GWC 2004, Proceedings*, pp. 234-241. © Masaryk University, Brno, 2003.

EuroWordNet-2, offering some suggestions to improve concept relatedness based on the formation of concepts in natural languages.

Many other works focus on building wordnets in other languages, or the creation of multilingual databases containing several wordnets. The former aspect will be addressed in chapter 3. In the next section, multilingual wordnet databases will be discussed in greater detail.

1.2 Multilingual wordnet databases

Many works have been written about the following projects. As this section is only meant to be introductory, illuminating only those aspects that are essential and could prove relevant for our project, each of these sub-sections provide only brief overviews on their topics.

1.2.1 *EuroWordNet*

EuroWordNet (EWN) is a multilingual database containing wordnets for eight European languages: English, Dutch, German, French, Spanish, Italian, Czech and Estonian. As mentioned before, they are all interconnected through an Inter-Lingual-Index (ILI), which is an unstructured list of concepts in English. This list is based on WordNet 1.5, but adapted for a more efficient mapping (see *EuroWordNet Final Report*, Vossen 1999:3).

EWN was built from 1996 to 1999 through two contracts. The first one was concerned with the languages English, Dutch, Spanish and Italian, as well as the EWN database development. The second contract, called EuroWordNet-2 or EWN2, covered the development of additional wordnets for German, French, Czech and Estonian (1999:7).

The approach of building the database was guided by the following objectives, quoting from Vossen (1999):

- to create a multilingual database;
- to maintain language-specific relations in the wordnets;

- to achieve maximal compatibility across the different resources;
- to build the wordnets relatively independently (re)-using existing resources;

(1999:11)

Additionally, Vossen states that they make a distinction between "language-specific modules and a separate language-independent module". The language-specific modules are "autonomous" systems of "language-internal relations between synsets".

The ILI is, as one could infer, part of the language-independent module. Each synset in each wordnet has at least one equivalence relation with an ILI record. The ILI is unstructured because of its unique role of being the mediator between different wordnets, which is its only purpose (1999:11). The ILI is, however, structured by two language-independent ontologies linked to ILI records. Quoting from Vossen (1999:11), they are:

- the Top Concept ontology, which is a hierarchy of language-independent concepts, reflecting important semantic distinctions, e.g. Object and Substance, Location, Dynamic and Static;
- a hierarchy of domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Traffic, Road-Traffic, Air-Traffic, Sports, Hospital, Restaurant;

The author also states that these ontologies are linked to the ILI records and can therefore be applied indirectly to all language-specific concepts linked to these records, as well as their inherited concepts. The main purpose of the Top Ontology is "to provide a common framework for the most important concepts in all the wordnets". The domain labels "can be used directly in information retrieval (and also in language-learning tools and dictionary publishing) to group concepts in a different way" (1999:11-12), aiding in conceptual disambiguation.

In order to achieve maximum overlap between the different wordnets, the notion of Base Concepts was introduced. This is explained on the official web site of the Global

WordNet Association[38]. According to the site, these concepts are the ones that are supposed to play the most important role in multiple wordnets. Two criteria for choosing these concepts are stated, namely that they should have a high position in the semantic hierarchy, and that they should have many relations to other concepts. They are the "fundamental building blocks for establishing the relations in a wordnet and give information about the dominant lexicalization patterns in languages".

The Base Concepts have been extended in the BalkaNet project, which will be discussed in section 1.2.3. A subset of 4689 concepts, called the Common Base Concepts[39], are freely available on the web site of the Global WordNet Association[37] and should be considered a useful, if not necessary, starting point for building a wordnet that is to be linked to the ILI.

Two different approaches to constructing wordnets have been introduced in this project, the so-called "merge" and "expand" models. In *EuroWordNet Tools and Resources Report* (Vossen et al 1998), they are defined as follows (quoting):

- **Merge model**: the selection is done in a local resource and the synsets and their language-internal relations are first developed separately, after which the equivalence relations are generated to WordNet1.5. This approach is followed for the Dutch and Italian wordnets.
- **Expand model**: the selection is done in WordNet1.5 and the WordNet1.5 synsets are translated (using bilingual dictionaries) into equivalent synsets in the other language. The wordnet relations are taken over and where necessary adapted to EuroWordNet. Possibly, monolingual resources are used to verify the wordnet relations on non-English synsets. This approach is followed for the Spanish wordnet.

  (1998:6)

Examples of these methodologies will be investigated in more depth in chapter 3.

[38] http://www.globalwordnet.org/gwa/gwa_base_concepts.htm
[39] These are concepts that function as Base Concepts in at least two languages, see http://www.globalwordnet.org/gwa/gwa_base_concepts.htm.

The particular design of EWN has several advantages, such as multilingual information retrieval, interlingual comparison, independent development and expansion of additional wordnets and the sharing of the language independent modules (1999:13).

In EWN, some changes have been made to the language-internal relations, such as the labelling of certain relations (for example, as "disjunctive" or "conjunctive"), as well as the addition of more relations that cross parts of speech (such as the relations between "die" and "dead", or "adorn" and "adornment") (1999:14).

EWN has many applications, some of which are mentioned by Vossen in the same work:

> In addition to the use for (cross-language) information retrieval, there are many other applications that can directly benefit from the multilingual semantic resources: information-acquisition tools, authoring-tools, language-learning tools, translation-tools, summarizers. Furthermore, they can be used for enabling technologies such as: word-sense-disambiguation, improve speech recognition, spelling checkers, parsers, language generation tools.
> (1999:15)

The completion of EWN has seen a great many results. Already, many wordnets were being developed at the time according to the specification of this database. Furthermore, says Vossen, "The results of EuroWordNet have also been used and integrated in EAGLES[40], Simple[41] and the ANSI[42] committee for standardized ontologies" (1999:16).

---

[40] EAGLES: Expert Advisory Group on Language Engineering Standards. It is an initiative of the European Commission (EC), with the aim of providing standards for large-scale language resources, ways of manipulating them, as well as assessing and evaluation these resources, tools and products. (Source: Official website - http://www.ilc.cnr.it/EAGLES/home.html)

[41] "SIMPLE is a large project (…) in the framework of the Language Engineering programme, and represents an innovative attempt to develop wide-coverage semantic lexicons for twelve languages (…) with a harmonised common model that encodes structure semantic types and semantic (subcategorization) frames." (SIMPLE: A General Framework for the Development of Multilingual Lexicons; www.lingfil.uu.se/personal/viberg/Lenci.pdf.)

[42] ANSI: American National Standards Institute. "The American National Standards Institute (ANSI) coordinates the development and use of voluntary consensus standards in the United States and

EWN has spawned other projects working within its model. Quoting from the article *WordNet, EuroWordNet and Global WordNet* by Vossen (2002:13-14), they are:

- The EUROTERM project extends the EuroWordNet database with specialized terminology (Stamou et al. 2002, http://www.ceid.upatras.gr/Euroterm/).
- The BALKANET project extends the database with more European languages: Czech, Romanian, Greek, Turkish, Bulgarian, and Serbian (Stamou et al. 2002, http://www.ceid.upatras.gr/Balkanet/)
- The MEANING project extends the database with sense-tagged corpora extracted from the WWW and word-sense-disambiguation modules (Rigau et al. 2002, http://www.lsi.upc.es/~nlp/projects/meaning.html).

Because of a lack of time and resources, some desired aspects could not be attended to, according to the author. He summarises these in the form of a list of requests made to the developers:

> - include adjectives and adverbs;
> - further improve the equivalence mapping,
> - further develop the ILI,
> - include multi-words and expressions,
> - integrate EuroWordNet and Parole.
> (1999:18)

Some of these issues are being attended to in the context of the Global WordNet Assocation. This is discussed in section 1.3.

1.2.2 *MultiWordNet*

MultiWordNet (MWN) is a multilingual wordnet project, at first concerning the construction of an Italian WordNet that is strictly aligned with the PWN,

---

represents the needs and views of U.S. stakeholders in standardization forums around the globe." (From: Official web site, http://www.ansi.org)

"The development of such a generic top-ontology is the aim of the ANSI-committee on Ontology Standards. Their Reference Ontology includes about 3,000 general concepts taken from a variety of existing resources." (*Higher Level Ontologies*, www.ilc.cnr.it/EAGLES96/rep2/node23.html)

distinguishing itself from the EWN specification in a variety of ways. In EWN, the approach has mainly been the separate building of wordnets, followed with the task of finding correspondences between them (Pianta et al 2002:1, quoting Vossen (1998)[43]), with the Spanish WordNet being one of the exceptions. The latter invented the so-called "expand model", already mentioned in the section on EuroWordNet, 1.2.1. Following in its footsteps, the MWN approach has been to try to build all Italian synsets in correspondence with the English synsets and import the relations between them as is (2002:1). The authors quote Vossen[44] (1996) in saying that "the expand approach seems less complex and guarantees the highest degree of compatibility across different wordnets." However, they also quote him as saying that this model depends heavily on the lexical and conceptual structure of one of the involved languages (2002:1).

One significant difference between the MWN model and that of EWN – and by extension the Spanish WordNet - is the absence of the Inter-Lingual-Index (ILI). Even so, there exist automatic procedures that can "speed up both the construction of corresponding synsets and the detection of divergences between PWN and the wordnet being built" (2002:2). Two main procedures are used in the building process: the "Assign-procedure", which consists of the construction of Italian synsets weighed against lists of possible PWN equivalent synsets, and the "detection of *lexical gaps* (LG-procedure), which are cases when a lexical concept of a language is expressed through a free combination of words in another language" (2002:2).

The interface to the database is, at the time of writing, accessible on-line on their web site, http://multiwordnet.itc.it/english/home.php. According to the site, the Spanish, Hebrew and Romanian wordnets are also compatible with the MWN specification, but, even though they are accessible through the on-line interface, they are not part of the MWN distribution.

[43] Vossen, P., ed., (1998) *EuroWordNet: A multilingual database with lexical semantic networks.* Kluwer Academic, Dordrecht.
[44] Vossen, P. (1996) *Right or wrong: combining lexical resources in the EuroWordNet project*. Proceedings of Euralex-96 International Congress.

Also at the time of writing, the MultiWordNet project is, according to their official web site[45], still ongoing.

1.2.3 *BalkaNet*

BalkaNet is a multilingual wordnet database containing wordnets for the languages Czech, Greek, Bulgarian, Romanian, Turkish and Serbian. Its initial aim was to extend the set of languages represented in EWN with five languages from the Balkan area, namely Bulgarian, Greek, Romanian, Serbian and Turkish, with the intent to achieve a wider cross-lingual coverage than in EWN[46] (Tufiş et al 2004:11). The main set of goals agreed upon from the beginning of the project is, quoting from Tufiş et al (2004:13-14):

- developing at least 8000 synsets per new language-specific wordnet, commonly selected so that even with this small size, the wordnets should be useful in real applications;
- ensuring maximal interlingual overlap among the BalkaNet wordnets and compatibility with the wordnets developed in the EWN project;
- building free software tools for the efficient management and exploitation of the multilingual semantic lexicon;
- development of application demonstrators such as Word Sense Disambiguation (WSD), intelligent document indexing, cross-lingual Information Retrieval (CLIR), etc.

One difference between BalkaNet and EWN is the nature of the ILI: Whereas the ILI of EWN was based on the PWN 1.5, the developers of BalkaNet have, although it is still defined in the same way as in EWN, updated it so that it is based on the PWN 2.0. This does not pose a problem with compatibility, according to the authors, since "more than 90% of the mappings among different versions of PWN (1.5, 1.6, 1.7.2, 2.0) are done automatically" (Tufiş et al 2004:14).

The selection of a common set of 8000 concepts to be used in all concerned wordnets was to "warrant a significant overlap among the BalkaNet wordnets" and "ensures a

[45] http://multiwordnet.itc.it/english/home.php
[46] Note that they do not claim to have actually achieved this goal.

satisfactory degree of conceptual intersection across wordnets and facilitates the cross-lingual evaluation and comparison of the monolingual repositories" (2004:17). This set consists of a selection of EWN base concepts, as well as others that were chosen according to certain criteria, such as their being lexicalised in at least two of the languages involved (2004:18).

As stated before, one main aim of the project was to develop applications for it, such as a system for word sense disambiguation and intelligent document indexing (see 2004:14). Another goal, also mentioned here, was the building of free software tools for various purposes related to the project. One of them is VisDic, developed at the Masaryk University of Brno, Czech Republic, which is a multilingual viewer and editor and can be downloaded, at the time of writing, at http://nlp.fi.muni.cz/projects/visdic/. A newer wordnet and dictionary development tool, by the same developers, is called DEBVisDic and is "built on the recently developed platform for client-server XML databases, called DEB II" (Horák et al 2006:1). This will also be discussed in chapter 2.

In the conclusion of their paper, Tufiş et al mention that "[a]ll the six wordnets have a significant cross-lingual coverage via PWN" and that "[b]y adopting the EWN methodology, the BalkaNet wordnets are extending the pool of aligned wordnets to an unprecedented semantic network for 15 European languages", also noting that they are "immediately usable in various applications" (2004:34-35). With its strong application-oriented approach and free available software tools, BalkaNet has certainly made a positive contribution to the wordnet scene.

1.3 The Global Wordnet Association

Building on the results of EWN and the PWN, this non-profit organization aims, according to the official web site, to maintain, standardise and interlink wordnets for all languages in the world, "likewise preparing the ground for the development of a world-wide multilingual database with wordnets" (*Background Document*, http://www.globalwordnet.org/gwa/gwa_background.htm). One main reason for the forming of the Association was the fact that other groups wanted to develop wordnets of their own and needed some common framework and guidance in this regard (Vossen 2002:14).

At the time of writing, Dr Christiane Fellbaum of Princeton University, who worked on the PWN, and editor of the book *WordNet: An Electronic Lexical Database* (1998), is the President of the Board, while Prof Piek Vossen, project coordinator of EWN, is the Vice-President.

Also at the time of writing, there are in total 47 different wordnets registered at the GWA that are either already built or in the process of being built, according to the official web site[47].

Their main aims are stated next, quoting from the web site (http://www.globalwordnet.org/gwa/gwa_background.htm):

- standardize the specification of lexical semantic relations, the notion of a synset and degrees of polysemy
- standardize the Inter-Lingual-Index for interlinking the wordnets of different languages
- extend the specification to include all Parts-of-Speech and multiword expressions
- develop a common XML representation for wordnet data
- prepare the development of sense-tagged corpora in all the linked languages
- sharing and transferring of data, software and specifications across wordnets for different languages
- the development of guidelines and methodologies for building wordnets in new languages
- the development of explicit criteria and definitions for verifying the relations in any language
- the development of consistency checking, comparison and evaluation modules

These attempts at standardization are clearly very important for future wordnet builders, in order for them to reach maximum compatibility, if they are all built according to these specifications. Also, especially important for building a resource-

[47] http://www.globalwordnet.org/gwa/wordnet_table.htm

scarce wordnet for a language such as Afrikaans, is the abovementioned development of guidelines and methodologies. Personal contact with Prof Vossen, as well as the wide-spread availability of papers on methodologies on the Internet, has made a tremendous contribution towards the fulfillment of this work.

The GWA has established a framework for building, evaluating, maintaining, improving, interlinking and applying wordnets, of which the author believes one should fully take advantage.

CHAPTER 2: Investigation of the feasibility of the project: A detailed look at all relevant factors

Following is a presentation of the relevant factors concerning the feasibility of the proposed project. First of all, we provide some specific background information.

2.1 Team background

All people who are directly involved in the project are employees of the Centre for Text Technology (CTexT) at the Potchefstroom campus of the North West University, South Africa. In recent years, they have developed various core technologies for Afrikaans and other South African languages, such as spell checkers, hyphenators, part of speech taggers and lemmatisers, as well as computer assisted language learning (CALL) software for four of these languages. At the time of writing, they are also in the process of developing a single language learning programme for all 11 official languages of South Africa, which is also web enabled, including the incorporation of WAP technology. They have also developed corpora, which will be discussed in the next section, and have converted an Afrikaans thesaurus into machine-readable format. In 2007, they are taking part in an enormous machine translation project, also involving all the official languages, as well as, among others, the construction of a wordnet for Afrikaans.

Apart from the advantages stated in section 1.1.2, we have chosen the wordnet model because of the following reasons:

- It is economically feasible.
- It can be constructed with relatively limited resources.
- It can take a relatively short time to develop it into a useful application and resource.
- There is a lack of lexical databases and other electronic resources available for South African languages.
- It provides a useful environment to put our relatively newly developed core technologies to use. For example, we can use our lemmatiser to contribute to

the same user interface functionality as the PWN, allowing the user to type in a word of any inflectional form and redirecting it to the appropriate selection of synsets.

- As mentioned before, the potential to link it to almost any other wordnet via an Inter-Lingual-Index promises enormous advantages for cross-linguistic research and application.
- We believe that we have the expertise required to build it and that it would constitute a big leap for natural language processing in South Africa, which at the moment is still in its infancy when compared to the level of advancement present in Europe today.
- Future development of wordnets for those other (minority) South African languages whose resources are even lesser developed will empower those languages and their speakers.

In chapter 5 thoughts on possible applications of our wordnet are presented.

2.2 Overview of the project

The project team has received a grant for working on a wordnet for one year. Personal communication with Prof Piek Vossen from the Free University in Amsterdam, The Netherlands, Vice-Chair of the Global WordNet Assocation and also a confirmed colloborator in the project, has suggested that, given the right resources, tools and staff, it should be possible to construct a small core wordnet in this time.

The following experts are appointed for 2007. Most of them should be available for direct or supporting work on the project:

Management	Head
	Portfolio Manager
Support	Financial Manager
	Communications Officer
	Office Assistant
Development	Program Manager (x3)
	Computational Linguist (x3)

	Project Manager (x3)
	Project Assistant (x4)
	Developer: System
	Developer: Core Technologies
	Developer: Data
Research	Senior Researcher

Figure 3: List of employees potentially available for work on the project

Both the core technology and the data developers are trained linguists. The latter is also trained as a lexicographer and professional translator. Of course, the staff members are not appointed exclusively for work on the wordnet, but other projects as well. Some of them are temporarily employed, but the periods could be extended. All staff members are fluent in both Afrikaans and English.

Each person also has a personal computer running on the Microsoft Windows XP operating system with access to the local Intranet and the Internet. Additionally, the Centre has a computer lab with 16 computers, most of which also run on Windows XP, one currently on the new Windows Vista and a few on Windows 2000 and 98 for testing purposes.

It should be obvious that building a wordnet requires some lexicographic resources. The following monolingual sources will be available for use:

- De Stadler, L.G. (with the assistance of Amanda de Stadler) 1994. *Groot tesourus van Afrikaans*. Halfweghuis: Southern Book Publishers (*Great thesaurus of Afrikaans*)
- Harteveld, P. (with the assistance of L.G. De Stadler and D.C. Hauptfleisch) 1993. *Woordkeusegids: 'n Kerntesourus Van Afrikaans*. Halfweghuis: Southern Book Publishers. (*Word Choice Guide: A Core Thesaurus of Afrikaans*)
- Odendal, F.F., Schoones, P.C., Swanepoel, C.J., Du Toit, S.J. and Booysen, C.M. 1994. *Handwoordeboek van die Afrikaanse Taal*. Midrand: Perskor. (*Desk Dictionary of the Afrikaans Language*)

It is highly probable that the PWN is going to be used as a source as well. At the time of writing, WordNet 2.1 is available on the official web site of the Princeton University (http://wordnet.princeton.edu/). In order to make any use of it at all, we need at least one bidirectional bilingual source, namely an Afrikaans-English English-Afrikaans dictionary. One standard bilingual dictionary is available for use on the project:

- Eksteen, Louis Cornelius. 1997. *Groot Woordeboek: Afrikaans-Engels, Engels-Afrikaans*. Kaapstad: Pharos. (*Major Dictionary*: *Afrikaans-English, English-Afrikaans*).

There is also a bilingual dictionary of prepositions:

- Taljaard, P.J. 1987. *Voorsetselwoordeboek Met Engelse Vertalings Asook Enkele Bywoorde*. Pretoria: De Jager-HAUM. (*Dictionary of Prepositions with English Translations Including Some Adverbs*)

Additionally, a few lists of bilingual terms are available:

- List of statistical terms in Afrikaans with English translations, definitions in both languages, as well as contextual information.
- Microsoft Core Terms: A list of computer-related terms in English with Afrikaans translations and English definitions.
- Dorfling, Johan. 2004. *Brugwoordelys (Engels – Afrikaans) en omskrywings van Afrikaanse begrippe*. (*Word List of Bridge*[48] *Terms (English – Afrikaans) and descriptions of Afrikaans concepts*)

Also, there are monolingual word lists, which may not prove very valuable, but which may at least be checked to confirm the existence and correct spelling of a word, as well as to place it in a specific domain. They are:

[48] The card game.

- Bouerswerktuie (*Construction Tools*)
- Sterrekundeterme (*Astronomy Terms*)

Of course, we shall use an official spelling guide for all spelling related issues:

- *Afrikaanse Woordelys En Spelreëls* / saamgestel Deur Die Taalkommissie Van Die Suid-Afrikaanse Akademie Vir Wetenskap En Kuns. Suid-Afrikaanse Akademie vir Wetenskap en Kuns. Taalkommissie. Kaapstad : Pharos, 2002. (*Afrikaans Word List and Spelling Rules / compiled By The Language Commission of the South African Academy for Science and Arts. South African Academy for Science and Arts. Lanuguage Commission. Cape Town: Harps, 2002.*)[49]

The current standard corpus used is the *Puk/Protea-Korpus Geskrewe Afrikaans* (*Puk/Protea Corpus of Written Afrikaans*) which consists of more than 2 million words. Recently, a parallel corpus called the *NWU Bible Corpus*, consisting of three languages - Afrikaans, English and Dutch - has also been developed. Additionally, there are two lists available, one for Afrikaans to English and the other one for English to Afrikaans, containing words with possible translations in the corpus, each one with a translation probability expressed as a value between 0 and 1.

Because of the relative small size of the *Puk/Protea* corpus, it is probably not wise to make any large-scale generalisations from the data. At the most, it could be used to support inclusion of a word in the lexicon by providing, in the case of monosemous (i.e. having only one sense) words, a relatively high frequency of occurrence, or to support an already determined fact, such as the validity of a perceived collocation that is indicated as such in a dictionary. Of course, the non-existence or low frequency of a word in the corpus could also suggest that it should not be included in the wordnet.

The Bible corpus must be considered unreliable, because many words are archaic and/or only occur in the context of the Bible. However, some word-translation pairs[50]

[49] Henceforth also referred to as *AWS*.
[50] They are strictly speaking, of course, not necessarily pairs, as a word may have more than one translation equivalent. By "pair" is meant the pairing of a word with a set of translation equivalents.

may be considered if they match with their respective translations in one or more bilingual dictionaries and if an Afrikaans word in such a pair also has a relatively high frequency in the *Puk/Protea* corpus.

With regard to tools, the one to be used for wordnet building and editing is the server-client based programme called DEBVisDic, which was already mentioned in section 1.2.3. Courtesy of the Faculty of Informatics of Masaryk University in Brno, Czech Republic, this tool was released, free of charge, shortly before the time of writing.

The only confirmed collaborator in the project is Prof Vossen as mentioned above. It is highly likely that a few more groups, probably local, are going to be involved, but, at the time of writing, this has not yet been verified.

Past experience has shown, according to Prof Vossen (personal communication), that it takes about a year to build a small core wordnet, using more or less the same group of experts that are at our disposal. Over the years, methodologies have been refined and currently, several automatic and semi-automatic methods exist for potential use in the wordnet construction process. It is therefore of the author's opinion that it is quite reasonable to believe that in our case, it should be possible to achieve the set outcomes. Of course, in order to enforce this belief, we need to take a deeper look into what is available and how it could be utilised.

2.3 Description of lexicographic resources

First of all, each lexicographic resource is inspected. At the time of writing, some works in electronic format are not available (*Handwoordeboek van die Afrikaanse Taal, Groot Woordeboek*); in those cases, we investigate the printed versions[51].

2.3.1 *Groot Tesourus van Afrikaans* (*Great Thesaurus of Afrikaans*)[52]

Available in the standard markup format XML, this is a monolingual thesaurus in the traditional sense, where seemingly monosemous broad concepts act as domain markers and other, usually more specific concepts, are grouped under the broader one.

[51] In section 2.1 it was mentioned that they are available. With this was meant that they are available for the wordnet project. At the time of writing, we are still in the process of obtaining these resources in their respective electronic formats.

[52] Henceforth also referred to as *GTA*.

This seems similar to hypernyms and hyponyms, but can cover more than one sense and word category. Also, a "broader concept" can have one or more entries in its list that could be considered synonyms of the concept itself. To give an example: "Noodsaak" (n. *necessity, need*; v. *force, compel, oblige, necessitate, entail, occasion*)[53] is the name of an article. The first section name is "noodsaaklik" (*necessary, essential, needful, imperative, prerequisite*), followed by 42 words that could be considered synonyms or near-synonyms. To demonstrate, the first ten are listed here, with English translations following in brackets:

- absoluut noodsaaklik (*absolutely necessary*)
- essensieel (*essential, vital*)
- onmisbaar (*indispensable, essential, vital, necessary*)
- dwingend (*compelling, compulsive, coercive*)
- nodig (*necessary, needful, proper, wanted, required, requisite*)
- broodnodig (*absolutely* (*highly*) *necessary, badly needed, essential*)
- dringend nodig (*urgently necessary*)
- allernodigs (*most necessary*)
- hoognodig (*very (highly) necessary, urgently needed, much needed*)
- ontontbeerlik (*indispensable, essential, imperative*)

It is clear that some of these words denote different grades of the same concept, such as "nodig", "broodnodig" and "dringend nodig".

Under the same article, "Noodsaak", there is another section name, "voorwaardelik" (*conditional, qualified, contingent, provisory*), but no synonyms are displayed. The next section name is called "noodsaak", with a small "n", and also includes a list of synonyms, containing phrasal units such as "Dis 'n moet" (*It is a must*).

[53] This and the next set of translations in section 2.2.1 and 2.2.2, except for "absoluut noodsaaklik", "hoognodig", "dringend nodig", "oondbak", "liefdadigheidsgawe", "gunsie" and "blikkie", for which there were no translations, are taken from:

Grobbelaar, Peter (ed.). *Afrikaans-Engelse Woordeboek / English-Afrikaans Dictionary*. Cape Town: The Reader's Digest Association South Africa.

One could spot a difference with the next list, where the article name is "Houer" (*container, vessel, carrier, dispenser*) and the sections under it have names such as "pot" (*pot, jar*) and "kastrol" (*stew-pot, stew-pan, saucepan*), which are different kinds of containers but co-hyponyms of each other. Under "pot", names such as "blik" (*tin, (tin) can, canister*), "blikkie" (*(small) can*) and "souspot" (*gravy-boat, sauce-boat*) appear, and under "kastrol" there are words such as "drukpot" (*pressure-cooker*), "oondbak" (*oven dish*) and "braaipan" (*fry(ing)-pan*). It is clear that these words, occurring on the same level, are not synonyms of each other but, just like their respective hyperonyms, mutual co-hyponyms.

It seems, therefore, that the existence of words occurring in the same list would suggest that they could appear on the same level of a hypernym/hyponym hierarchy, being either synonyms (or, for that matter, near-synonyms) or co-hyponyms of each other. This could be used as part of a method to assign confidence scores to translation candidates in the synset construction process, but only after lists have been verified to contain either synonym or co-hyponym candidates.

There is also a text file available which is another presentation of the content of the thesaurus. It provides, on each line, a list of synonyms. Inspecting some fragments reveals that they are indeed synonyms, but that other synonyms, or even the same word forms used in the same senses, are found in other lines. This is unfortunate, but in this case, the existence of two or more words in the same line would strongly suggest that they belong to the same synset.

2.3.2 *Woordkeusegids* (*Word Choice Guide*)[54]

This work, which is also a thesaurus, has a similar layout to the GTA. On the surface, it looks promising. Also ordered in a hierarchy, broader concepts contain lists of synonyms divided up by its senses. For example, under the word "aalmoes" (*alms, charity, dole, baksheesh*), the first sense is numbered "1", followed by synonymous words such as "liefdadigheidsgawe" (*gift of charity*), "geskenk" (*present, gift, offering, donation*) and "gawe" (*gift, present, donation*). A second sense is numbered "2" and contains the words "guns" (*favour, custom, goodwill, kindness, patronage,*

[54] Also abbreviated in this work as *WKG*.

support), "gunsie" (*little favour*) and "hulp" (*help, aid, assistance, support, succour, relief*), all of which can also be seen as synonyms of each other. It is clear that sense 1 and 2 are distinct, and that they could be of help in the synset construction process.

However, there is a problem: The different unrelated senses of homonyms are not distinguished, at least not in some cases. For example, the word "blik", which is a homonym, refers to a list containing both the words "visie" (*vision*) and "kan" (*can, jar, jug, mug, tankard, pitcher, pot*). This seems like a serious flaw, because, obviously, it is necessary to make those sense distinctions when building synsets. However, each word in a list and its head word can be considered synonym candidates.

2.3.3 *Handwoordeboek van die Afrikaanse Taal* (*Desk Dictionary of the Afrikaans Language*)[55]

Generally considered the leading standard dictionary for Afrikaans, it has a good coverage, with detailed but clear word entries and a comprehensive usage guide, including lists of general abbreviations and those used to indicate style, domain, etc. in the lexicon.

A typical entry is written in this form (here, and not in the dictionary, different parts are separated, for clarity, by three spaces):

lemma (pronunciation) **spelling variant** (inflectional information) label definitions: *usage*. [etimology] **list of compounds and/or derivational forms see also**

For multisyllabic lemmas, stress is also indicated, such as *ou'klip*. Homonyms are treated as separate entries, such as **baai**[1] and **baai**[2] while polysemous distinctions are numbered withing the same entry. These distinctions are not always limited to the same parts of speech. For example, *leer* as a noun (in the senses of *teaching, doctrine, theory, apprenticeship, doxy, gospel*) and *leer* as a verb (in the senses of *teach, instruct, indoctrinate, train, tutor, break in* (a horse)*, learn* (s.t. from a person)) all

[55] Also abbreviated in this work as *HAT*.

occur in the same entry. We do not expect this to be a problem in the electronic version.

The printed version of the dictionary does not always provide parts of speech or inflectional information, in order to save space. As this could pose a major problem for the wordnet, we assume that the electronic version does provide all relevant information. Unfortunately, at the time of writing, it is not yet in our possession.

A potential point of contention is the fact that different entries whose forms are identical with the exception of some diacritical marks, such as *voel* (*feel*) and *voël* (*bird*), are classified as homonyms. This is, of course, useful for entering search lemmas in a browser – diacritical marks can be ignored, leading to search results where they are both present and absent. However, this can also lead to problems in automatic processing, if not handled carefully, where, for example, one of the said forms may not be recognized as belonging to a set of homonyms.

Here is an example of a relatively complete entry:

> **le'ë** b.nw en bw. Verboë vorm van *leeg: Leë houers, kanne, bottels.* s.nw. (-s) Iets wat leeg is; leë houer: Leës moet teruggestuur word.

The "b.nw. en bw." means "adjective and adverb". It is followed by the definition, which states that it is an inflected form of the word *leeg*. This is an irregular inflected form, which demonstrates that these are included as lemmas. The example follows after the colon. After the second square, a "s.nw." appears which is an abbreviation for "noun", followed by "(-s)", which is the morpheme to be added in its plural form. Once again, it is followed by a definition, with the example after the colon.

Entries can be as short as this:

> **leer'kontrak** Kontrak met bepalings omtrent leerjongens.

In this case, the information after the lemma is the definition.

Sometimes, synonyms are given as definitions, such as “Slag, hou, stoot” under “coup” (from the French). If these are extracted, they could be used to verify sets of synonyms obtained from translating PWN synsets.

Genus extraction is another option, based on the fact that many definitions take on the following form: *An X is a Y (that Z)* or ***X** (A) Y (that Z)*. Usually, Y could be seen as standing in a hypernymic relation to X. For example, a *bicycle* is a kind of *wheeled vehicle*, where *wheeled vehicle* is a hypernym of *bicycle*[56]. If Z exists, it usually contains additional information about Y, narrowing the potential semantic space that X can reference. However, it is a known fact that dictionaries are not very consistent with their methods of writing definitions. It stands to reason, then, that such extractions would probably not be very successful. One also needs to construct a parser beforehand in order to achieve this. Finally, the correct meaning of the genus still needs to be determined. In the expand model, extracting relations from dictionaries is not strictly necessary, as the semantic relations are taken over from PWN as is. As, from a preliminary viewpoint, it seems unlikely that we are going to use the merge methodology (more time and resources required, which we lack), the former seems like an attractive choice. Synonym extraction can also be considered, a possibility that we are going to investigate further in chapter 4.

Apart from the latter, the *HAT* can also be used to extract glosses, as well as lexical relations such as derivational forms, variants and words with similar meanings, via the *sien* (*see also*) link.

2.3.4 *Groot Woordeboek* (*Major Dictionary*)[57]

This is a standard bilingual dictionary with fair coverage, comprising 43 536 entries. It has a simple structure, where a basic entry is written in this form (SL = source language; TL = target language):

SL lemma, TL equivalents; SL usage, TL equivalent usage.

[56] Princeton WordNet 2.1.
[57] Also abbreviated in this work as *GW*.

Every multisyllabic lemma has a stress indication, as in the following example:

he'llo girl, (U.S.)

Note the optional label here indicating the geographical usage. Morpho-syntactic information is also provided, although the word category is not indicated, except to disambiguate homonyms with different parts of speech; the above example also demonstrates that morphological information is also not always available. We assume the reasons are that standard inflectional forms are left out in order to save space, indicating that this dictionary is meant for relatively advanced users, that is, those who can also control the second language to some degree. As with the *HAT*, we can only assume that this information will be complete in the electronic version.

Following standard practice in most standard desk dictionaries, homonyms are distinguished by separate entries. A useful feature is the fact that synonymous translation equivalents are separated by commas, while the borders between polysemous sets are indicated by semi-colons. Note for example the next entries, the first from the English-Afrikaans part and the second from the Afrikaans-English part:

hell, hel; duisternis; speelhol, dobbelnes (…)
hel[2,] (w) **(ge-)**, lean, incline, slope, slant; list; lean over; (…)

In the first entry, there are three sets of translation equivalents:

1. hel
2. duisternis
3. speelhol, dobbelnes

In the second entry, there are another three sets:

1. lean, incline, slope, slant
2. list
3. lean over

Clearly, these look like preliminary synsets, which is certainly a useful feature to have in a dictionary when building a wordnet. However, in the translation sets, sense distinctions are not made. The *hell* entry fails to mention which sense of "hel" is meant as a translation equivalent. One does not need knowledge of Afrikaans to notice that "hell" does not fit in with the words "lean", "incline", slope", etc. Therefore, the human user can deduce that *hel*[2] is not the sense meant in the entry *hell*. In dealing with automatic translation and synset construction, a computer needs more information than that, though. Fortunately, a set of very useful techniques exists to deal with this problem while building a wordnet. This is further explained in chapter 3.

Here are two more entries to demonstrate additional features of a typical entry. The second entry is truncated here to indicate the continuation of the displayed pattern:

> **foon**[1]**,** (s) **(-s)** = **fon**.
> **foon**[2]**,** (s) **(fone)**, phone; **~boek,** phone book, telephone directory; **~flerrie**, callgirl;…

In the first entry, the (s) indicates a noun, the **(-s)** displays the plural suffix and the "=" is used to inform the user that the following word and the lemma are variants of each other. For more information, the user then has to look up *fon*.

The second entry **(fone)** is an irregular plural form. After the only translation equivalent, *phone*, a list of possible combinations is shown. Note that even here, synonymous concepts are separated by commas (*phone book, telephone directory*).

Accompanying this entry is a text box with special information about the lemma in the source language (Afrikaans), mentioning, in this case, that *foon* is a shortened form of the word *telefoon*.

Finally, the lexicon is supplemented by four lists of abbreviations. Two are lists of general abbreviations used in either of the languages with translation equivalents in the other language in abbreviated form, while the other two are lists of editorial

abbreviations that are used in the entries to indicate information about the lemmas or their equivalents.

The author believes that this lexical resource will surely prove useful in extracting translations for the wordnet.

2.3.5 *Voorsetselwoordeboek* (*Dictionary of Prepositions*)

In this work, the lemmas are mostly single words, under each of which a list of grammatical collocations or idioms, which may contain other content words, with translations is to be found. The lemmas themselves are not translated. They are also not necessarily in their base forms, but are included as they are used in their respective collocations. Here is an example of an entry:

daad

by:	die daad *by* die woord voeg / put one's words *into* action
op:	*op* die daad / *at* once, there and then, immediately
:	iem *op* heter daad betrap / catch sb red-handed, catch sb *in* the act
tot:	*tot* die daad oorgaan / proceed *to* action
van:	'n man *van* die daad / a man *of* action

Figurative expressions are indicated as such, for example in the following case:

Daantjie

tot:	iem *tot by* oom Daantjie *in* die kalwerhoek opdreun / fight sb *to* the limit
vir:	sê groete *vir* oom Daantjie [*fig*] / cut it *out*, drop it [*ie s subject*]

Note that the lemma here is a proper noun, and even though the examples can be considered idiomatic, the prepositions used in these constructions are highlighted.

Clearly, this resource may be more difficult to use in the wordnet construction, also since this is not a collection of word-level senses – sense distinction occurs on the phrasal level. However, these are excellent for demonstrating a word in context, and may serve as a source for glosses or gloss translations.

2.4 Description of tools

2.4.1 *DEBVisDic*

DEBVisDic is a development tool for wordnets and other lexical resources, designed at the Faculty of Informatics, Masaryk University, Brno, Czech Republic[58]. The precursor of this computer application is VisDic, which was used with success in the BalkaNet project for specifically editing wordnets. DEBVisDic can be used to develop not only wordnets, but "various types of dictionaries, i.e. monolingual, translational, thesauri and multilingually linked wordnet-like databases" (Horák et al 2005:1).

However, the same authors mention that the latter has some disadvantages. Quoting from this work, they are:

- It is not based on the client/server architecture.
- It does not allow to associate various attributes with literals[59] and handle the links between them. It can work with links only between synsets which is a limiting feature for enriching wordnets with various sorts of information (…) (2005:1)

These disadvantages have prompted the developers to design "a more universal dictionary writing system that could be exploited in various lexicographic applications to build large lexical databases" (2005:1) which was called *Dictionary Editor and Browser*[60] (DEB), the final version being known as DEBII. The authors mention a few projects that were being implemented as clients at the time of their writing the article, within this platform. To quote:

- DEBVisDic – new version of wordnet editor and browser (…)

---

[58] The official web site is http://nlp.fi.muni.cz.

[59] With "literals" is meant single entries within synsets, which can be a word or phrase. The "links between them" refers to the lexical relations as opposed to synset relations in a wordnet.

[60] Horák et al (2006:1-2) refer in this regard to:
Pavel Smrž and Martin Povolný. 2003. Deb – dictionary editing and browsing. In *Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003*, pages 49-55, Budapest, Hungary.

- DEBDICT – general dictionary browser. This simple DEB client demonstrates several basic functions of the system:
    - multilingual user interface (English, Czech, other [languages] can be easily added)
    - queries to several XML dictionaries (with different underlying structure) with the result displayed with the use of XSLT transformations
    - connection to a morphological analyzer
    - connection to an external website (Google, Answers.com)
    - connection to a geographical information system (display of geographical data directly on their positions within a cartographic map)
- Czech Onomastic Dictionary – newly prepared dictionary of Czech proper names and their origins
- PRALED – new Prague Lexical Database of Czech

(2005:3)

This platform is defined as following a "strict client-server architecture" (2005:2), leading to the authors call a "homogeneity of the data structure and presentation" (2005:3), such that any changes made on the server regarding data presentation is automatically reflected in the client software.

Simply stated, DEBVisDic is a reimplementation of VisDic for this particular platform, with new features added for supporting wordnet construction (2004:3). Some useful assets mentioned here are:

- All the data are stored on the server, the latter of which also provides a considerable amount of functionality, in contrast with a simple functionality on the client's side.
- This is beneficial for team cooperation – all changes made in the server are seen by everybody, as mentioned above.
- In this manner, data flaws can also be corrected in one step.

- It is also possible for the client to work offline "in a degraded manner", without a connection to the server.

(2004:3)

The authors also state that DEBVisDic has "high modularity and configurability" (2004:4-5).

One slight disadvantage also mentioned here (2004:3) is that for a fully functional application, an operating server and network connection is necessary.

Horák et al provide a list of features that were also present with Visdic. Quoting:

- multiple views of multiple wordnets
- freely defined text views
- synset editing
- hypero-hyponymic tree
- query result lists
- plain XML view of a synset
- synchronization
- inter-dictionary linking
- tree browsing
- consistency checks
- journaling
- user configuration

(2004:3-4)

Some added features also mentioned here are (quoting):

- connection to a morphological analyzer (for languages, where it is available)
- connection to language corpora, including Word Sketches[61] statistics
- access to any electronic dictionaries stored within the DEB server

[61] Word Sketch is a software tool used for the extraction and presentation of significant collocations for lexicographic purposes.

- searching for literals within encyclopedic web sites
- and many others

(2004:4)

It is the author's belief that, especially for building a wordnet for a resource-scarce language such as Afrikaans, with limited time and finances, using DEBVisDic as a central tool in the construction process is an absolute necessity.

2.5 Preliminary conclusion

Since we have not investigated some successful methodologies yet, the above so far only suggests, from a preliminary viewpoint, that we have ample time and resources at our disposal for building a small core wordnet for Afrikaans in one year. It is still premature to roughly estimate a number of synsets, let alone a specification of the relations or structure involved, but more details follow in chapter 4.

In the next chapter, we investigate a few main approaches to building wordnets.

> Taking this into consideration, as well as what we have discussed in this chapter, we reach a conclusion in chapter 4 and recommended a methodology of our own, which may or may not be based on one or more already existing ones.

Chapter 3: Investigation into various approaches to wordnet construction

The first version of the Princeton WordNet came out almost 20 years before the time of writing this work and the project is still continuing. Although today it is still the most extensive wordnet by far, its construction was painfully slow and most occurred by hand. It is therefore quite obvious that, for us, this is not the way to go.

As mentioned before, in recent years there have been many improvements in wordnet construction methodologies, simply because so many wordnets have been built and because there has been a world-wide sharing of expert information in this regard. Already during the EuroWordNet project, a distinction was made between the so-called merge and expand methodologies, which was mentioned in section 1.2.2. Sometimes a mixed approach is followed, as with the Portuguese wordnet[62].

Following is an investigation into a few successful and well-documented methodologies used. We give an example of a merge methodology (Dutch), the original expand methodology (Spanish), a different expand methodology with new automatic procedures (MultiWordNet) and another expand methodology with some different automatic procedures (a Romanian wordnet of nouns based on the Romanian wordnet in BalkaNet). Details that are too specific, such as most numerical statistics, or of processes that are not relevant, such as the conversion of data from the Dutch database Vlis to the EuroWordNet structure, are not mentioned. However, as much detail as possible will be provided in our recommended methodology discussed in chapter 4.

3.1 Example of a merge methodology: The Dutch WordNet

All the information on this methodology comes from an official document on this project, *The Dutch WordNet* (1999) by Piek Vossen, Laura Bloksma and Paul Boersma.

---

[62] See: *Portuguese WordNet: general architecture and internal semantic relations*. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502002000300008&lng=es&nrm=iso&tlng=en.

As was mentioned before, the merge methodology consists of first building the wordnet and then aligning it with a chosen target wordnet. The construction of the Dutch WordNet (DWN) was part of the EuroWordNet project. The developers were fortunate in having at their disposal a rich collection of lexicographic resources. Quoting from Vossen et al (1999:5), this is the list:

- Vlis: the content of a lexical database provided by Van Dale[63];
- Vde: the Van Dale Dutch-English dictionary (Martin and Tops 1986);
- Ved: the Van Dale English-Dutch dictionary (Martin and Tops 1989);
- WordNet 1.5 (Fellbaum 1998);
- Celex Dutch and English lexicons with basic morpho-syntactic information and corpus frequency information on word forms and lemmas;
- a list of new Dutch spellings (according to the spelling convention of 1997) extracted from a Dutch-Russian dictionary build[sic] at the University of Amsterdam;

A great advantage for the DWN is that the Vlis database, containing 63,962 nouns and 8,822 verbs, already contained explicit semantic and morphological relations between nouns, verbs and adjectives (1999:5).

The Van Dale Dutch-English and English-Dutch dictionaries, constituting 90,925 and 89,428 entries respectively, were developed for native speakers of Dutch, containing more information on English words than the reverse. The authors call it a "rather basic resource" (1999:7-8), although it does have extensive coverage.

Moving on to the methodology, the authors mention that they followed the merge approach, because a structured database already existed, and that they were "particularly interested in differences of the lexicalisation patterns in Dutch and other languages." (1999:8)

As was already discussed in section 1.2.1 under EuroWordNet, a top-down approach was followed where each wordnet built their respective cores around a common set of

[63] Van Dale is a leading dictionary publisher in The Netherlands.

Base Concepts in order to ensure a degree of compatibility across the different wordnets. These concepts were mostly based on PWN 1.5 concepts and together with the rest of the PWN they form the Inter-Lingual-Index. As the authors mention here, they play a major role in at least two wordnets, typically having a generally high position in the hierarchy and connected via relations to relatively many other concepts (1999:8-9).

3.1.1 *Extracting the translation equivalents*

Translations have mostly been done using the bilingual dictionary. Synsets were translated by mapping the Van Dale database with the bilingual Dutch-English dictionary and the translations to WordNet 1.5 (1999:12). This means that the set of Dutch words occurring in both the database and the dictionary were translated using the dictionary, using the PWN synsets in which these translations occur as the set of all possible translation equivalents. As could be expected, not all translations were possible in this process. Some other methods were used to increase the number:

- Dutch words that are directly taken over from English are omitted in the dictionary. These words were directly matched to the WordNet 1.5 entries in which these word forms occur. Some nouns were matched correctly, which were then intersected with the synsets without translation for an improved translation percentage.
- Some further improvements were made by reversing the English-Dutch dictionary in a Dutch-English dictionary.
- More synsets could be translated by "varying the use of hyphens and spaces in the translations." These characters are sometimes used inconsistently – the examples shown are *animal park, animal-park* and *animalpark*. Varying the forms in this manner in the translations enabled them to translate 338 more synsets.

  (1999:12-13)

In the translation process, naturally more than one translation is often possible. For words that have just one translation, it was assumed to be correct; for the rest, a set of heuristics was devised to decrease the number of possible translations per word.

1. The first heuristics consisted of checking if the additional information accompanying a Dutch word in the bilingual dictionary matches its equivalent in the Vlis database. This was done using string comparison. The information could consist of a definition or some morpho-syntactic features. The results are used to assign a weighting to the relevant translations.
2. The second heuristics consisted of a technique known as conceptual distance as defined by Agirre and Rigau[64] (1996), mentioned by the authors. There are two aspects to this process:
    a. The "distance between senses of multiple alternative translations of a single entry in the bilingual dictionary" (1999:14) is measured. The example given here is the Dutch word *orgel*, which can have two translations, *organ* or *keyboard*. Using conceptual distance, it is possible, for example, to determine the right sense of *organ* to be linked to *orgel* by selecting the sense which is closest in the hierarchy to *keyboard*.
    b. The "distance between each possible translation and the translations of hyponyms and hyperonyms of the Dutch word" (1999:14) is measured. The example given here is, once again, the Dutch word *orgel*, where one translation equivalent is *organ*, which could be a *musical instrument* or a *body part*. A hypernym of this particular sense of *orgel* (*muziekinstrument*) is linked to a hypernym of a particular sense of *organ*, namely *musical instrument*. Regarding their respective hyponyms (*hammond orgel* and *hammond organ*), the same analogy is applied. The distance between this particular sense of *organ* and *hammond organ*, as well as between *organ* and *musical instrument* is measured, the same is done with other senses of *organ*, resulting in the senses with the shortest distances as being the most likely translation candidates (see figure 4). The results were analysed and problematic cases, such as a large number of translations with equally good matches, are isolated and the particular entries manually translated. Words with many translations and relations have also been translated

[64] Agirre E. and Rigau G. 1996. Word Sense Disambiguation using Conceptual Density, in proceedings of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen. 1996.

by hand, namely verbs with more than 2 relations and more than 10 translations, as well as nouns with more than 10 relations and more than 10 translations. The conceptual distance technique is further discussed in the context of the Spanish methodology in section 3.2, where a formal algorithm is presented as well.

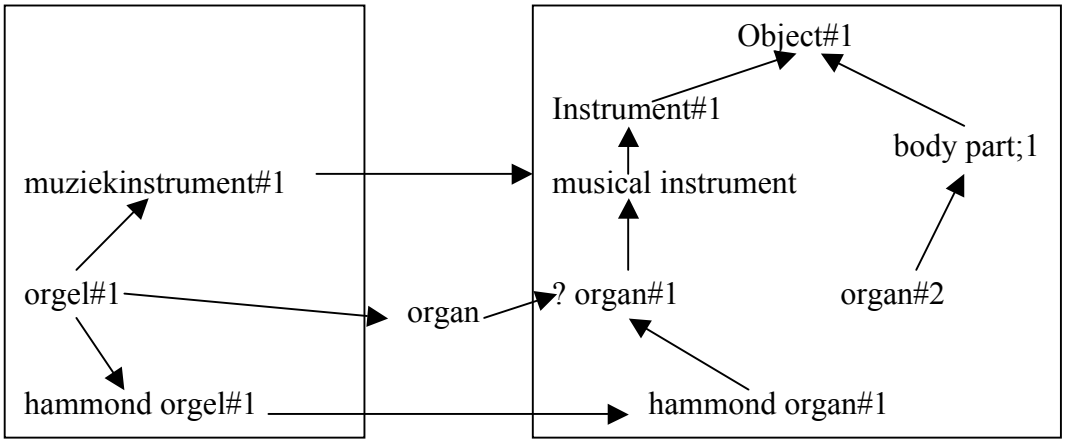

Figure 4: Modification of Figure 2, p.15, http://www.vossen.info/docs/1999/DutchWordNet.pdf (Selecting translations to WordNet1.5 by distance to the translated context in the Dutch wordnet)

3. Thirdly, WordNet 1.5 synsets were translated back to Dutch and the overlap with Dutch synsets is measured. Vossen et al (1999:16) provide an algorithm:
    a. take the possible candidate translations generated from the Dutch-English resource to WordNet 1.5;
    b. look up the target variants in the English-Dutch resource;
    c. increase the match:
        i. each time an English variant has a variant of the Dutch source synset as its translation;
        ii. if multiple variants of the Dutch source synset are given as the translation for a single English sense;
4. The fourth heuristic measures the "overlap in Top-Concepts inherited according to the Dutch hierarchy for the Dutch sense and according to the WordNet 1.5 hierarchy for the target translations" (1999:14). By this is meant the EWN Top Ontology, which was separately added to the Base Concepts in DWN and WordNet 1.5. These concepts "represent fundamental semantic

> features, such as Natural, Artifact, Dynamic, Static, Physical, Mental" (1999:17) and all hyponyms of those synsets that are associated with these features also inherit these features. Through this, the overlap in Top Concepts between Dutch senses and their possible set of translations can be measured. Obviously, if the overlap is greater, this indicates that the particular translation equivalent is a better candidate. The example given here is the Dutch word "hart" (heart as an organ) which inherits the features *Living* and *Part* from its hypernyms *orgaan, deel* and *iets*, senses 1, 2 and 1 respectively. Only sense 4 of the PWN synset in which "heart" appears have these features as well, suggesting that this synset should be linked to the Dutch equivalent in which "hart" (in the sense of the organ) appears.
> (1999:14-18)

Further on in Vossen et al (1999), the results are shown. The aforementioned techniques proved effective. Manual editing and translation were done for the remaining problems, such as lexical gaps.

3.1.2 *Construction of the core wordnet*

In selecting the Base Concepts (BCs) for Dutch, the builders of the DWN were in a unique situation in the sense that they had the Vlis database, already containing a hierarchy of concepts, from which to work. Some procedures were followed to arrive at the final set of BCs, the details of which we are not going to pursue further, except to say that more than 50% of these concepts are also Common Base Concepts which occur in other wordnets as well (1999:25).

The core wordnet is constructed around the selection of BCs. Quoting Vossen et al (1999:25), the following general procedure was followed to process the BCs:

1. select closely related terms and word meanings;
2. establish the major semantic classes and differentiation;
3. provide a hyperonym classification to the top of the hierarchy;
4. chart out hierarchical differences among the selected words;
5. establish the correct equivalence relations for the most important concepts;
6. establish any other language-internal relation in so far necessary;

The minimal contents of a core wordnet, including "all relevant relations to the local equivalents of the Base Concepts", are:

- hyperonyms;
- 1st level of hyponyms below the Base Concepts;
- equivalence relations;

(1999:9)

Whereas the Vlis database is not a closed system where all relations "are unified in a single tree or a small set of tops" (1999:22), this is the case for the Dutch wordnet, as stated by the authors (1999:6). It then follows that every concept that is added as part of the core wordnet must directly or indirectly be related to one or more BC.

To be able to make a list of the concepts that should be added, some techniques were followed:

1. Adding co-occurrence data from corpora. Research has shown that words that occur frequently in the same context tend to have similar meanings.
2. Closely related synsets in WordNet 1.5 and their hyponyms were translated into Dutch, sometimes revealing "other classes and categories that have not been thought of or have not been included in first Dutch set" (1999:25).
3. Definitions that are very similar or circular definitions suggest that their respective lemmas may be synonyms.
4. Words from sets generated such as in 3. were translated from Dutch into English and then back-translated. The set of back-translated words that were also in the original set "is used as a filter to select the correct meaning of the word."
   (1999:25-26)

This added list of concepts and their relations were then studied. Some concepts were found to be misrepresented in the hierarchy and had to be moved in another position. The developers also made other changes such as splitting and merging synsets, reallocating hyponyms after the major classes have been established, specifying

translations for major classes, and, where necessary, further enriching BCs or closely related concepts with non-hyponymy relations (1999:28).

3.1.3 *Extending the core wordnet to a complete Dutch wordnet*

In the same work as discussed above, the authors mention that the focus in the second building phase for the DWN was:

1. extending the core wordnet to the full size;
2. improving the overlap across the wordnets;
3. improving the quality of the equivalence relations;
4. verify the corpus frequency of the selection;
5. include regular lexicalization patterns;
6. adding new spelling variants;

(1999:29)

For the purpose of this work, we are only going to discuss point 1. The following steps are all mentioned in Vossen et al (1999:29-31).

1. First of all, those semantic features described by the Top Ontology that had relatively few concepts clustered under it were investigated and new concepts were added to balance out the Top Ontology coverage.
2. Lexical gaps were also investigated and synsets were added where possible. In some cases they were not genuine gaps – they were not included because of various possible reasons, such as differences in the hierarchy or the translation being a multi-word, a set of entries that was not included in the first phase.
3. The coverage of DWN senses when compared to the Dutch Parole lexicon were increased to 100% for all senses with a frequency above 100.
4. New spelling variants were added to the synsets.
5. All synsets with 10 or more relations, all direct hyponyms of concepts with 50 or more relations and all synsets with 1 or 2 automatically derived translations were included, as well as all hypernyms that were needed to classify these concepts.

6. Finally, some concepts were excluded, such as those with no translation or a low frequency.

3.1.4 *Conclusion*

It is clear that the merge methodology used to build the Dutch WordNet is complex and time-consuming, demanding a very well thought-out order of procedures and meticulous work. It is of the author's opinion that the developers of the Afrikaans wordnet should not use this approach for their project, considering relatively scarce resources and the lack of an existing database. However, some methods mentioned here could prove useful, such as the variation of hyphens and spaces in English words to improve translation coverage, as well as the method of applying the conceptual distance formula. Somewhat quicker and simpler is the expand methodology, which we are now going to discuss firstly by using the Spanish WordNet as an example.

3.2 Example of an expand methodology: The Spanish WordNet

It was already mentioned that the expand methodology entails deriving the target wordnet from a source wordnet, with the result that the former, at least originally, assumes the structure of the latter, as opposed to first building the target wordnet and then aligning it with the source wordnet, as is done using the merge methodology.

3.2.1 *Methodology*

The Spanish wordnet was the first instance of this approach. For the main description of the methodology, the cited work is *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets* by Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau and Horacio Rodríguez (1997). They discuss a few different methods that are combined for the automatic construction of a preliminary Spanish wordnet, in the context of EuroWordNet. Using the expand method, the resulting structure is identical to the Princeton WordNet, in this case version 1.5. This is only possible because of the close conceptual similarity between English and Spanish (1997:2).

The lexical resources used here are:

- Spanish/English and English/Spanish bilingual dictionaries[65].
- A large Spanish monolingual dictionary[66].
- English WordNet, version 1.5.

(1997:2)

The two bilingual dictionaries have been merged to create what the authors call a "homogeneous bilingual (Hbil)" (1997:2).

Distinguishing itself from the merge methodology, the first main step was not to construct an autonomous wordnet from monolingual sources, but to link the Spanish lexical entries to PWN synsets. Three kinds of methods are presented by the authors that are used in this process. Quoting from Atserias et al (1997:2), they are:

- Class methods: use as knowledge sources individual entries coming from bilinguals and WN synsets.
- Structural methods: take profit of the WN structure.
- Conceptual Distance methods: makes use of knowledge relative to meaning closeness between lexical concepts.

Samples of the results were taken and inspected, yielding a confidence score (CS), which depends on the correctness of these results. Five different possible results can be obtained from an automatic translation link (1999:2-3): correct, fully incorrect, links to a hyponym of the correct synset, links to a hypernym of the correct synset, and links to near synonyms that could be considered correct.

The ways in which translation links were made are also inspected; more specifically, how many links from how many words to how many words. The different cases are:

- For monosemous words:
    - One Spanish word is translated by one English word. The reverse is also true.

[65] Diccionario Vox/Harraps Esencial Español/Inglés – Inglés/Español Biblograf S.A. Barcelona 1992
[66] DGILE: Diccionario General Illustrado de la Lengua Española – Vox – M. Alvar (ed) Biblograf. S.A. Barcelona 1987

    - One Spanish word is translated by more than one English word. Each English word has only this Spanish word as its translation.
    - More than one Spanish word is translated by one English word only. The English word has all these Spanish words as its possible translations.
    - More than one Spanish word has more than one English translation. These English words also have more than one Spanish translation.
- The criteria for polysemous words are identical, except that the English words are polysemous.
- For variants: If a WN1.5 synset contains two or more variants that have only one translation to the same Spanish word, this word is linked to the particular synset.
- For English words whose entries have field identifiers in the dictionary: If two of them occur in the same synset, they are linked to all of their Spanish translations.
  (1997:3-4)

The structural methods make use of the PWN structure to disambiguate translation equivalents. For every Spanish entry, all English translations were generated using Hbil, followed by finding as much common information between these English words in WordNet 1.5 as possible (1997:4). For every different case, another action was followed. Here is the algorithm for the whole process, quoting from Atserias et al (1997:4):

- Intersection criterion
  Conditions: All EWs share at least one common synset in WordNet. Link: SW[67] is linked to all common synsets of its translations.
- Parent criterion
  Conditions: A synset of an EW is a direct parent of synsets corresponding to the rest of EWs. Link: The SW is linked to all hyponym synsets.
- Brother criterion

[67] EW = English word; SW = Spanish word

Conditions: All EWs have synsets which are brothers respecting to a common parent. Link: The SW is linked to all co-hyponym synsets.

- Distant hyperonymy criterion
  Conditions: A synset of an EW is a distant hypernym of synsets of the rest of the EWs. Link: The Spanish Word is linked to the lower-level (hyponym) synsets.

At first, these links may seem inappropriate and possibly even incorrect, but the results show, after the deletion of unnecessary repeating entries, impressive scores of between 81,39% and 94,4% correct links. Moreover, the larger the list of translations, the higher the precision and the lower the amount of incorrect links (1997:4).

Regarding conceptual distance methods, the authors first cite Rada et al (1989)[68] in defining the notion of conceptual distance (CD): "(…) the length of the shortest path that connects the concepts in a hierarchical semantic net." (1997:4). Atserias et al also mention Agirre et al (1994)[69], who devised a formula for calculating this distance, which is shown here:

$$dist(w_1,w_2) = \min_{\substack{c_{1i} \in w_1 \\ c_{2j} \in w_2}} \sum_{c_k \in path(c_{1i}, c_{2j})} \frac{1}{depth(c_k)}$$

Figure 5: Conceptual distance formula, as shown in Atserias et al (1997)

The *w*'s indicate words and the *c*'s are synsets in which those words occur (1997:4). For example, c_1 indicates the synset in which w_1 occurs; if there is more than one synset, another number follows c_1, hence the variable in the formula (eg. c_{1i}). The two points of reference are the group of synsets containing w_1 and those containing w_2, i.e. c_{1i} and c_{2j}. The c_k denotes any synset occurring in any given path connecting c_{1i} and c_{2j}, while the result of the fraction on the right hand side is added to itself for every

[68] (Rada *et al*. 89) R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development an Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17-30, 1989.
[69] (Agirre *et al*. 94) E. Agirre, X. Arregi, X. Artola, A. Díaz de Ilarraza, and K. Sarasola. Conceptual Distance and Automatic Spelling Correction. In *Proceedings of the workshop on Computational Linguistics for Speech and Handwriting Recognition*, Leeds, UK, 1994.

occurrence of c_k for a given path. [*M*]*in* indicates that the smallest of all those possible values are selected and accorded to the variable *dist(w1,w2)*. Finally, $depth(c_k)$ refers to the depth of c_k in the hierarchy. The result of the formula can be used to find the concepts representing w_1 and w_2 that are the closest to each other.

The formula has been applied using three different methods (1997:5)

- Assuming the notion that two words appearing in the same definition in a dictionary are co-occurrent (mentioning Wilks et al (1993)[70]), a list of more than 300,000 co-occurrence pairs have been extracted from DGILE, the Spanish monolingual dictionary used. The "affinity between these pairs was measured by means of the Association Ratio (AR)[71]" (1997:5) after which the CD formula was applied to all these pairs.
- The formula was also computed on the headword and genus term of more than 92,000 nominal definitions occurring in the DGILE.
- Finally, the developers derived a small lexicon of just over 3,000 translation equivalents from Spanish to English of the bilingual dictionary.

Afterwards, they tested the results by measuring how often the formula connected the actual closest concepts. A table in (1997:6) reveals that the first method scored only 56%, the second 61% and the last 75%.

Taking all the previous methods combined (class, structural and conceptual distance), those synsets produced by methods that were more than 85% accurate were collected and combined to produce a preliminary version of the Spanish WordNet, "containing 10,982 connections (1,777 polysemous) among 7,131 synsets and 8,396 Spanish nouns with an overall CS of 87,4%" (1997:5). Then the results of the discarded methods were combined and their intersections calculated, resulting in an additional

[70] (Wilks *et al*. 93) Y. Wilks, D. Fass, C. Guo, J. McDonal, T. Plate, and B. Slator. Providing Machine Tractable Dictionary Tools. In Pustejowsky J., editor, *Semantics and the Lexicon*, pages 341-401. Kluwer Academic Publishers, Dordrecht, 1993.

[71] The Association Ratio is a score "which can be used to test the strength of the bond between the two variables of a contingency table", which is "close to the concept of mutual information" (Daille 1995:32, citing Church and Hanks (1990)):

> [Church and Hanks, 1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. 16, nº 1, pp. 22-29.

group of synsets and links, of which those with a confidence score of over 85% were added to the first version of the wordnet.

The authors conclude their work by stating that the "approach seems to be extremely promising, attaching up to 75% of reachable Spanish nouns and 55% of reachable WN1.5 synsets" (1997:6). They also stress the fact that the methods combined produced more accurate results than those from the individual ones. However, this was not applied to verbs and adjectives, which were more difficult to process and hence translated manually[72].

3.2.2 *Conclusion*

The class and structural methods apparently work well for any set of translation equivalents, including those that are not already divided up into sense groups. Their decisions are not well explained here but yield good results. The Conceptual Distance formula is an innovative and useful method to help choose one synset as an equivalence link over another. Also the fact that combining these methods leads to even better results is certainly worth considering. Next we consider a different expand method, as implemented by MultiWordNet.

3.3 Expand methodology 2: MultiWordNet

MultiWordNet (MWN) has been introduced in section 1.2.2. The work used as reference for explaining this particular approach is *MultiWordNet: Developing an aligned multilingual database* (2002) by Emanuele Pianta, Luisa Bentivogli and Christian Girardi.

As mentioned in section 1.2.2, MWN is a multilingual wordnet database constructed without the existence of the Inter-Lingual-Index as it is found in EuroWordNet. The first instantiation is an Italian wordnet which is strictly aligned to the Princeton WordNet[73]. As, according to the official web site, other previously developed wordnets that were also built using the expand method were added to the database, resulting in the ability to compare equivalent synsets in different languages, the PWN does actually function as a kind of ILI in MWN.

[72] According to the co-ordinator of EuroWordNet, Prof Piek Vossen (personal communication).
[73] The specific version of the PWN is not mentioned in the article.

In general, the approach followed here consists of two automatic procedures: the *Assign* and the *Lexical Gaps* (LG) procedures (Pianta et al 2002:2). Simply stated, the former attempts to build Italian synsets which are semantically equivalent to PWN synsets; where it is not possible, it is labelled an "English-to-Italian or an Italian-to-English lexical idiosyncrasy" (2002:2) and handled by the LG procedure.

The authors mention two possible strategies that can be followed building Italian synsets:

- For every English synset, obtain all Italian translation equivalents (TEs) for all the words in the synset. A set of TEs is considered the Italian synonymous synset of its PWN counterpart. If no TEs can be found, an English-to-Italian lexical idiosyncrasy has been found.
- For every Italian word *I*, obtain its TEs in English. Link *I* to all PWN synsets that contain at least one TE. Afterwards, for every PWN synset *S*, group together all Italian words that were linked to *S*. This group is the Italian synonymous synset of *S*.
  (2002:2)

Pianta et al speculate that the best alignment can most likely be achieved by combining both strategies, but that they have only followed the second one, also stating that Atserias et al (1997) have done the same.

3.3.1 *Assign procedure*

The Assign procedure helps the lexicographer by selecting, out of a set of all possible synsets, zero or more synsets that are more likely to match to a given Italian word. Each of these synsets are assigned a confidence score (CS). Only those with a CS above a certain percentage are considered. (2002:3)

In the bilingual dictionary, translation equivalents are grouped in senses by semi-colons. See section 2.3.4. for a demonstration of this in one of our lexicographic sources, the *Groot Woordeboek* (*Major Dictionary*). Thus, an entry can be divided up into its senses. A sense of an Italian word can therefore be matched up with a set of

candidate synsets in the PWN by selecting all those PWN synsets that contain at least one of the TEs occurring in the particular sense specification. The authors call these synsets the *set of candidates* or *CandSet* (2002:3). The CS of every synset in the CandSet is calculated by the application of certain linking rules. These rules are divided up into four groups. The successful application of a rule to a particular synset raises its CS. They are:

- generic probability
- back translation
- gloss matching
- synset intersection

(2002:3)

With the generic probability rule, it is assumed that only one synset in CandSet can be the right one to be matched to a particular Italian sense. Therefore, the bigger the CandSet, the lower the probability of each synset to be matched to the Italian sense, resulting in a lower CS. Conversely, if there is only one candidate synset, it is highly probable that it is the right one, resulting in a very high CS (2002:3).

The back translation rule is applied in the following way: If an Italian sense is linked to a PWN synset, the CS score is raised if one or more synonyms of that synset have the Italian word as a back translation. The example given here is the Italian word "puntura", translated as "sting", in the sense of referring to insects. This word occurs in four PWN synsets. One of these synsets also contain the word "bite", which can also be translated back to "puntura", increasing the likelihood that this particular synset is the correct one (2002:4).

Gloss matching compares Italian gloss information with their PWN counterparts. Quoting from Pianta et al (2002:4), this may contain one or more of the following:

- a semantic field specification (e.g. **"sclerosis *n*** *(Med)* sclerosi", where "*Med*" means medicine)
- a synonym (e.g. **"reason 1. *n* a.** *(motive, cause)* ragione,…")

- a hypernym (e.g. **"sole *n*** *(fish)* sogliola)
- a specification of the context of use (e.g. **"handle 1. *n* …** (*of knife*) manico, impugnatura; (*of door, drawer*) maniglia;…")

The existence of domain labels for the PWN makes it possible to match entries containing them. If an Italian sense and one or more of the synsets in CandSet contain labels that refer to the same domain, the confidence scores for those synsets are raised. (2002:4)

Words or phrases in Italian glosses are translated and the TEs are matched with the words in CandSet synsets. In general, the more they match, the higher the CS, but the more ambiguous the words involved, the lower the strength of the increase because of a higher chance of an incorrectly chosen sense. (2002:4)

This process is extended by making use of the fact that a gloss often contains a genus word (i.e. hypernym) of the word that it defines. Firstly, one tries to match the Italian word with the hypernym of its TE. Secondly, the same Italian word is compared to an English word from the gloss of a hypernym of the candidate synset. Any matches here raise the CS, but not as much as in the case of direct matching of Italian glosses and PWN synset words. (2002:4)

Additionally, extra gloss information expressed in specific types of constructions can be matched with their counterparts in PWN. For example, one definition of "fold" in PWN has the specification "of the skin", which can be matched to the Italian "della pelle", which means the same thing and is also applied to one of the Italian senses which is a TE of "fold (of the skin". (2002:4)

Synset intersection uses the fact that the words in translation groups (TGRs) are synonyms of each other. Synsets containing more TEs than others, for a particular Italian sense, can be considered better candidates. This, of course, results in a better CS for these synsets. The given example is the Italian word "pilastro" which is translated in its metaphorical sense as "pillar, mainstay". The authors state that "pillar" belongs to 5 PWN synsets, while "mainstay" belongs to 3. However, only one

synset includes both of these words. According to this rule, this makes it a more likely candidate than the other synsets. (2004:5)

The Assign procedure was evaluated using a number of nouns from the Collins dictionary. After applying these rules, a number of candidates with a CS of over a certain threshold[74] was selected and proposed to their lexicographers for assessment. The results were, quoting from Pianta et al (2004:5):

- 70% precision: The ratio between the number of candidates accepted by the lexicographers and the number of candidates proposed by the algorithm.
- 63% recall: The ratio between the number of candidates accepted by the lexicographers and the number of word senses listed in the Collins dictionary.

3.3.2 *Lexical Gaps procedure*

The authors group the previously mentioned "idiosyncrasies" that can arise as part of the translation process into two kinds. Quoting from Pianta et al (2002:5), they are:

- *lexical gaps*: a language expresses through a lexical unit what the other language expresses with a free combination of words (…)
- *denotation differences*: the TE of a source language exists but it is more general or more specific. In the former case the TE is a sort of cross-linguistic hypernym of the source language word and in the latter case it is a cross-linguistic hyponym (…)

The examples given here are "bell" that is translated in Italian to "campanello" which means "small/electric bell", as well as "campana" ("church bell"), and also the word referring to the bell found on a cat ("sonaglio"). These are all examples of more specific TEs.

The definition of "free combination" is also made clear here when the authors state that they consider idioms and restricted collocations as lexical units which can be

[74] The percentage of this threshold is not given here, just that the selection amounted to 89% of the senses listed in the dictionary.

considered potential TEs or synonyms, whereas any other combination of words is considered a genuine "free combination".

To handle the LG procedure, the developers invented a method for identifying lexical gaps semi-automatically, relying on the fact that dictionaries sometimes label a group of words as being an idiom or restricted collocation, as well as the fact that these three groups "exhibit certain structural regularities" (2002:5-6). The authors refer to Bentivogli et al (2000)[75] for more information in this regard.

The LG procedure classifies all TGRs as lexical units, lexical gaps or TGRs that need to be manually checked to be classified as being either lexical units or lexical gaps. The results of this procedure seem impressive, with a mere 10,6% of TGRs in the English-to-Italian section that needed to be checked manually and only 7,0% percent for Italian-to-English. Interestingly, only 1,0% and 0,9% of TGRs were classified as lexical gaps, respectively (2002:6).

Keeping in mind that the Italian wordnet in MWN is strictly aligned to the PWN but can also be allowed to represent any lexical idiosyncrasies (2002:10), the authors point out that Italian-to-English lexical gaps refer to a set of Italian synsets of which no English synsets exist; therefore, the Italian synsets must be added manually. The converse refer to English synsets for which no Italian equivalents exist, therefore they can be excluded (2002:6).

3.3.3 *The data model*

The MWN model makes a distinction between language-specific and common data, assuming that semantic relations are common and lexical relations are language-specific. The database (DB) containing the Princeton and Italian wordnets is therefore divided into three separate modules: A COMMON-DB, an ENGLISH-DB and an ITALIAN-DB. Synsets that are linked cross-linguistically are referred to as "multisynsets" (2002:6), the relations of which are described by the COMMON DB. Each of the modules also contains add-ons, which can add or overwrite existing data. For example, semantic relations that occur only in Italian can be described in the add-

[75] Bentivogli L., Pianta E. and Pianesi F. (2000) *Coping with lexical gaps when building aligned multilingual wordnets.* Proceedings of LREC 2000, Athens, Greece, pp. 993-997.

on for the ITALIAN-DB. Any lexical-specific idiosyncrasies are also encoded in the language-specific add-ons, as well as additional lexicographic information.

3.3.4 *Conclusion*

In their paper, Pianta et al explain their methodology very clearly. It seems well attuned to using bilingual dictionaries where the translation equivalents are divided up into sense groups. Our bilingual dictionary, the *Groot Woordeboek* (*Major Dictionary*), fulfills this criterion. We do not have to follow the MWN data model; the Assign and LG procedure can be applied to a set of Base Concepts that includes WordNet Domains and the Top Ontology, around which a core wordnet can be built using the same methods. These concepts are a set PWN synsets, which includes relations between them. See the discussion of EuroWordNet in section 1.2.1 for more information.

Lacking, however, in this work is some explanation into how the confidence scores were calculated. Getting this right is obviously of great importance.

From a preliminary viewpoint, the author believes that one could combine these methods with the Conceptual Distance formula, which was introduced in section 3.2.

Next, another expand method is investigated: the automatic construction of a Romanian wordnet that contains only nouns.

3.4 Expand methodology 3: A Romanian wordnet of nouns

In their work *Automatic Building of Wordnets* (2005), Barbu and Mititelu propose a method for automatically building a wordnet that is strictly aligned with a source wordnet. The instantiation of this method is a Romanian wordnet of nouns expanded from the Princeton WordNet. It consists of two phases: Firstly, they generate the synsets for the target language (TL) and map it onto source language (SL) synsets. Secondly, they use a method to identify and import salient relations from the PWN. The results are evaluated against an existing wordnet, the Romanian WordNet built as part of the BalkaNet project. (2005:99)

For the selection of synsets to be implemented, the developers followed two criteria (2005:100):

- Every selected synset should have at least one semantic relation with another synset.
- The result must be evaluated using a "gold standard", which in this case is the Romanian Wordnet (RoWN) from the BalkaNet project.

A group of synsets fulfilling the first criterion was chosen from the RoWN, which also includes the upper level concepts of the PWN. Projected against PWN 2.0 synsets, this comprises 9716 synsets containing 19624 literals (2005:101).

The developers used two resources: an "in-house" dictionary built from many sources, including a bilingual section, as well as the Romanian Explanatory Dictionary, which is a monolingual source. (2005:101)

Further on, they also state that ideally, "every sense of a word should be unambiguously identified by a set of connections; it should have a unique position in the semantic net." (2005:101). This, they say, is unfortunately not the case in the PWN, where some synsets have precisely the same connections to other word senses.

Following these statements, the authors present a series of heuristics for deriving and mapping the Romanian synsets onto PWN synsets. They summarise the ideas behind it in three points:

- increasing the number of relations in the source wordnet to ensure a unique position for every word sense, using an external resource that is linked to the wordnet, such as WordNet Domains;
- deriving useful relations between the words in the TL, using, among others, corpora, monolingual dictionaries and already classified sets of documents;
- taking profit of the structures built through these procedures during the mapping stage.

(2005:101)

3.4.1 *First heuristic rule*

This rule describes the automatic construction of a target synset using a very simple two-step decision tree:

1. If at least one word in the source synset is monosemous, the target synset contains all the TEs of the monosemous word.
2. If all the words in the source synset are polysemous, the target synset consists of the intersection of all the TEs of all the words in the source synsets, that is, only those TEs that are linked to all the source synset words.

(2005:101-102)

In this specific case, the results were compared to the already existing Romanian WordNet. Barbu and Mititelu distinguish five different possible cases. Quoting, they are:

1. The synsets are equal (this case will be labelled as Identical).
2. The generated synset has all literals of the correct synset and some more. (Over-generation).
3. The generated synset and the golden one have some literals in common and some different (Overlap).
4. The generated synset literals form a proper subset of the golden synset (Under-generation).
5. The generated synset have no literals in common with the correct one (Disjoint).

(2005:102)

The cases of Identical and Under-generation are regarded as successes, while the others are errors. The reason that Under-generation is correct is because it does not contain any wrong literals; it is just not fully complete. (2005:102)

The results are impressive: Out of 8493 synsets mapped, only 2% were errors and the vast majority (7983) being correct (Identical) mappings. However, the authors, at least

partly, ascribe the success of the results to the quality of the dictionary that they used. (2005:102).

3.4.2 *Second heuristic rule*

This rule exploits a few common facts:

- The hypernymy relation can be viewed as an IS-A relation.
- Hypernyms and their hyponyms carry some common information.
- The amount of information that is common to the hypernym and hyponym increases going down in the hierarchy.

(2005:102)

The rule determines if two different source synsets that stand in a hypernym/hyponym relation to each other, can be mapped to a single target synset. In other words, the SL makes a distinction between concepts standing in this relation that is not made in the TL. In section 3.3.2, where the Lexical Gaps procedure in the MultiWordNet model was explained, this is illustrated by the English word "bell" that can only be translated to Italian to a more specific term. One of the TEs is "campana" which means "church bell". In PWN, "church bell" occurs in a different synset, which is a hyponym of "bell". This rule ensures that, for example, "campana" would link to both "bell" and "church bell", and vice versa. For Romanian, another example is given by the authors, although these do not stand in a hypernymy/hyponymy relation, but only demonstrates the general problem: The terms "mister" and "sir" are represented in two different synsets in the PWN, but they are both translated in Romanian by the word "domn". They state that "it would be artificial to create two distinct synsets for the word *domn*, as they are not different, not even in what their connotations are concerned" (2005:100).

Firstly, two source synsets that stand in a hypernymy/hyponymy relation to each other are selected. Then the second part of the first heuristic rule is applied to both of these synsets. That is, for each synset, the intersection of all the TEs of their respective literals is calculated. This results in two candidate synsets in the TL, one for each SL synset. The next step is to calculate the intersection of the two TL synsets. If this new

set is not empty, this is a strong suggestion that the TL has the same lexicalization for the two distinct concepts in the SL. In this case, the created synset is assigned to both SL synsets. (2005:102).

The authors foresee a complication in this process. If there exist, for example, an SL synset A, which is a hypernym of synset B, which is a hypernym of synset C, and synset A and B are linked to TL synset S, while synset B and C are connected to TL synset T: Choose the TL synset with the deeper level, that is, synset T. The reason is that, in this case, one needs to "perform a clean-up procedure and choose the assignment that maximizes the sum of depth level of the two synsets" (2005:102).

Results show that, for the Romanian case study, a very low number of synsets were mapped this way (10%). The authors state that this is because not many common translations between hypernyms and hyponyms were found. (2005:102). However, for the 1028 synsets that were mapped, 35% were errors (Table 2, 2005:104).

3.4.3 *Third heuristic rule*

Here, WordNet Domains are used to help construct TL synsets and map them to SL synsets. As stated earlier, the RoWN is used as a golden standard to evaluate results. However, this wordnet was aligned to PWN 2.0, while the Domain Labels were assigned to version 1.6. Therefore, the developers of this new Romanian wordnet performed a mapping between PWN 1.6 and PWN 2.0 (2005:102-103). Because of the fact that some new synsets in version 2.0 did not receive a domain label, some steps were followed to try and increase the coverage:

1. If a direct hypernym of an unlabelled synset has a domain assigned to it, the unlabelled synset will receive its domain. The same applies when a hyponym of an unlabelled synset has a domain assigned to it.
2. If a holonym of an unlabelled synset has a domain assigned to it, the unlabelled synset will receive its domain. The same applies when a meronym of an unlabelled synset has a domain assigned to it.
3. If, after this, a synset still has no domain label, it receives the default "factotum" label.

(2005:103)

The next step is labelling every word in the bilingual dictionary with its domain label. In the English section, the entries automatically receive them from the synsets. For the Romanian section, two methods were followed (2005:103)

1. Documents from web dictionaries were downloaded. The selection is such that their categories match the WordNet Domains. Then the documents were processed in the following ways:
    a. The words in the documents were Part of Speech tagged and lemmatised, after which the nouns were selected as possible terms that could describe their respective documents.
    b. The next step consisted of using the so-called χ^2 statistic, which "checks if there is a relationship between being in a certain group and a characteristic that we want to study" (2005:103). Using the following formula, they measured the dependency between a term *t* and a category *c*:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

Figure 6: The formula for the χ^2 statistic

Quoting, the terms in the formula are explained as follows (2005:103):

- A is the number of times t and c co-occur
- B is the number of times t occurs without c
- C is the number of times c occurs without t
- D is the number of times neither c nor t occurs
- N is the total number of documents

For every category, the χ^2 was calculated between that category and the noun terms of the documents. For choosing the best terms for a certain category, they used another formula (*m* is the number of categories):

$$\chi^2 \max(t) = \max_{i=1}^{m}(X^2(t, c_i))$$

Figure 7: Formula for choosing the best terms from an χ^2 statistic

2. The developers also used labels from dictionaries to assign domains to synsets by manually mapping them to the Domain Labels. If unsuccessful, they are labelled "factotum".

All of the previous steps result in all Romanian and English words having a domain label assigned to it. Now, when one constructs a TL synset from an SL synset, a specific algorithm is followed. Firstly, one has to keep in mind a few points. The authors provide a series of statements in mathematical notation, which we shall present here in text:

1. An SL synset has one or more domains assigned to it.
2. The literals in this synset have each, in the modified bilingual dictionary, one or more domains assigned to them, including but not necessarily limited to the domain or domains assigned to the particular synset.
3. The TEs of a word from the SL synset have each one or more domains assigned to them, which may or may not match the domain(s) of the SL word.

(2005:103)

The TL synset is constructed as follows:

1. Extract all TEs from all the words in the SL synset.
2. Collect those TEs whose domains match the domain(s) of the SL synset. More specifically, these TE domains are either:
    a. the same as the domain of the SL synset;
    b. or subsumes (includes) the domain of the SL synset in the domain labels hierarchy;
    c. or is subsumed by the domain of the SL synset in the domain labels hierarchy

3. Perform an intersection on these selected TEs as in the second step in the heuristic rule number 1 (section 3.4.1). That is, extract those TEs that are linked to every word in the SL synset.

(2005:103)

The results were again evaluated and once again they very good: Out of 7520 synsets mapped, only 9 percent were errors and the vast majority of matches (6831) perfect (2005:105).

3.4.4 *Fourth heuristic rule*

Here the authors present a complicated procedure which compares glosses from the TL and SL to each other, generating TL synsets based on criteria of similarity between the glosses of their individual words with those of the words in the SL synset. Once again, some preprocessing had to be done first, in the form of automatically lemmatising and tagging all glosses of the synsets in the SL, doing the same with the definitions of the words that are translations of SL words, after which the lemmatised nouns of the glosses were selected as features representing them (2005:103-104).

All definitions were thus represented by a set of nouns. The TL sets were automatically translated, resulting in a group consisting of all possible SL definitions in this noun set format, the latter called a "vector" by the authors (2005:104). They provide a formula for calculating the number of SL vectors:

- The definition of a TL word is represented by the following: [rw_1, rw_2, ... rw_p] where *rw* denotes a word, in this case a noun, in Romanian.
- The number of SL vectors generated are: N = n_d * t_{w1} * t_{w2} * ... * $t_{wp,}$ where n_d is the number of definitions of the target word in the monolingual dictionary and t_{wk} , with k = 1..p, is the number of translations that the noun w_k has in the bilingual dictionary.

(2005:104)

The set of SL vectors for a given TL word is represented by the term R_{Gloss}, using the following formula: $R_{Gloss} = \{T_1, T_2 \ldots T_n\}$, where *T* represents definitions for the TL word. S_v is used to denote the vector of the SL synset gloss (2005:104)

The next step is to generate the TL synset, using the following steps:

1. Generate all TEs for all the words in an SL synset.
2. For every TE, compute the similarity between S_v and its R_{Gloss}. This is done in two steps:
    a. A binary representation is given to the vectors in S_v and R_{Gloss}. The number of positions in a vector is equal to the number of distinct words existent in the S_v and in all vectors of R_{Gloss}. A 1 in the vector indicates that a word is present and a 0 means that a word is absent from the vector.
3. For every T_i vector in R_{Gloss} the product is calculated. The formula given is:

$$S_v \bullet T_i = \sum_{j=1..m} s_j * t_j .$$

Figure 8: Formula for computing the product of the source language synset vector and the vectors of the target language definitions

This means that every value in S_v is multiplied with every value in T_i and the results added to each other. The *s* is an SL word, the *t* a TL word and *m* is the number of positions in a vector. Then, if there exists at least one T_i such that $S_v \bullet T_i >= 2$ the $\max(S_v \bullet T_i)$ is computed and the word added to the TL synset. (2005:104)

The results show that there is a high number of incomplete synsets, which the authors ascribe to "the low agreement between the glosses in Romanian and English" (2005:104).

3.4.5 *Combining the results*

For evaluating the pros and cons of each rule, the developers "devised a set of meta-rules". Some examples are given (2005:104):

- Automatically select synsets generated from the first heuristic rule, since the quality of the dictionary is high and the probability of failing is low.
- A synset obtained from the other rules will replace a synset obtained from the first heuristic if:
    - it is obtained independently using the heuristics 3 and 2, or
    - by using heuristics 3 and 4.
- Else it will only be selected if:
    - it is obtained by the heuristics number 3 and
    - the ambiguity of its members is at most equal to 2.

These rules may seem confusing, but the results shown for the combined effort (2005:105) are very impressive: 98 percent (9610) of all synsets are mapped, of which only 9 percent are erroneous.

3.4.6 *Importing the relations*

Since the wordnet is conceptually strictly aligned with the PWN, the developers assumed that conceptual relations can be imported without any changes. The only lexical relation present was the antonym relation, of which they concluded, after some study, to be also safe to transfer as is. Next the authors describe the algorithm that they followed (2005:105):

1. Conditions:
    a. Two source synsets S_1 and S_2 are linked by a semantic relation R in the SL wordnet.
    b. T_1 and T_2 are the corresponding aligned synsets in the TL wordnet.

   Action:
    a. T_1 and T_2 are linked by the relation R.
3. Conditions:
    a. As in 1, but there are intervening synsets between S_1 and S_2.

b. R is declared as transitive (R+, unlimited number of compositions, e.g. hypernym) or partly transitive (R*k* with *k* a user-specialized maximum number of compositions, larger than the number of intervening synsets between S_1 and S_2)[76]. For example, all holonymy relations were defined as partly transitive (k=3).

Action:

a. T_1 and T_2 are linked by the relation R.

4. Conditions:
   a. As in 1, but there are intervening synsets between S_1 and S_2.
   b. Condition 3b is not met, i.e. R is not declared transitive or partly transitive[77].

   Action:
   a. T1 and T2 are not linked by the relation R.

Finally, the authors mention, among others, the work of Atserias et al (1997) and contrast this with their approach by stating some characteristics (2005:105):

- It gives an automatic evaluation of the results by automatically comparing them with a manually built wordnet.
- They explicitly state the assumptions of their approach.
- Their approach is the first to use an external resource, Wordnet Domains, in the process of automatic wordnet construction[78].
- They obtained a version of RoWN containing 9610 synsets and 11969 relations with 91% accuracy.

3.4.7 *Conclusion*

This methodology certainly appears quite useful. But first, one has to keep a few things in mind:

---

[76] Defining a relation as transitive means that all the semantic information is passed on through the relation. For example, the hyponym relation between "vehicle" and "car" is transitive because all the semantic information belonging to "vehicle" is also present in "car". A relation is partly transitive when not all semantic information is transferred, as with a holonymy relation.

[77] This could imply a relation such as "attribute". See figure 1 in section 1.1.1.

[78] This is not true, since WordNet Domains were used in both the Dutch WordNet and MultiWordNet for sense distinction. At least some of the steps involved in both cases were automatic. See sections 3.1.1 and 3.3.1, respectively.

- They had an already existing gold standard against which to compare their results (RoWN); we do not. The project team has to construct one themselves.
- They had a high quality bilingual dictionary; the status of ours is still to be determined.
- Their approach has only been proven to work for Romanian nouns.

Additionally, some processes involved, especially those mentioned in heuristics 3 and 4, are complicated and probably quite time-consuming. Others seem relatively simple and failsafe, such as heuristics rule number 1 and their algorithm for the import of relations. Their methods of combining the rules are not explained clearly and logically.

We have to consider to which methods we should give attention and which we should rather leave alone. For example, downloading on-line documents for feature extraction is probably not a good idea, as there is relatively little (structured and/or classified) on-line material available in Afrikaans. We also do not have a quality parser at our disposal for parsing glosses, etc.

The author believes that our approach should exploit all potentially successful methods in all the discussed methodologies and be more hesitant to use those for which it is less certain that they would yield good results. The logic behind our approach will be discussed in the next chapter.

Chapter 4: Statement of criteria, selection and presentation of a methodology for constructing the Afrikaans wordnet

4.1 Criteria for selecting an optimal methodology for the construction of the Afrikaans wordnet

In the last chapter we have presented four different approaches to building wordnets, each of which turned out to be successful for their respective projects. Of course, there are many factors that determine the success of building a wordnet. Some of them are obvious, and include:

- the quality of the lexicographic resources
- time and budget
- the expertise of the people working on it and their ability to work together as a team
- availability and successful application of automatic procedures
- the facilitation and successful application of manual and some automatic procedures through tools such as DEBVisDic
- the conceptual closeness of the two languages involved, in the case of the expand methodology
- proper consultation and collaboration with experts
- etc, etc.

Another obvious fact is that there are many proven methods to choose from, and that to combine them all would not be feasible. Therefore, some selection has to be made.

Deciding on a proper work schedule is complicated by two opposing goals:

1. producing a quality wordnet
2. keeping to the limited time frame and budget

Therefore, the approach to be followed must be guided by two priorities:

1. effectiveness, i.e. it must produce good, accurate results
2. simplicity, i.e. it must not be overly complicated and time-consuming

It is of the author's opinion that the first priority is more important than the second. This means that it is better to produce a small but high quality wordnet by the deadline than to have a bigger but sloppier result, although, of course, bigger is preferred, as long as the quality does not suffer because of it.

The obvious conclusion that we can draw is that, as far as possible, automatic procedures would be preferred, as long as they are accurate to an acceptable degree.

Two more factors to be considered are that the wordnet as a source must be reusable and compatible. Using it in isolation only is defeating the purpose of the project. As stated in section 2.1, a long-term goal is to expand it to other languages and be part of a potential worldwide linking of wordnets through an Inter-Lingual-Index. Therefore, as little changes as possible must be made to source files, such as the PWN structure, lemmas, synsets, domains, ontology, etc. Source IDs and relations must, as far as possible, be taken over and as many links made as possible. It must be able to be used within a variety of applications. The developers must always aim to better and expand it continuously.

Keeping all of the above in mind, it is clearly of great importance to make decisions regarding construction methods that not only speed up these processes, but also make sense in the long term.

Next, we present a basic step-by-step series of phases in the wordnet construction process, after which each one will be clarified in detail and recommendations made.

4.2 Building the wordnet: A step-by-step plan

Because none of the lexicographic sources were available in electronic format at the time of writing, none of the methods presented here could be evaluated. However, guidelines are given for the testing and evaluation of all the procedures involved.

These major phases, based on the approaches followed in the methodologies discussed in chapter 3, are distinguished in the construction process:

1. Selecting the methodological type: Merge or expand. We have predetermined that we are going to follow the expand methodology.
2. Selecting the source wordnet.
3. Selecting the lexicographic sources and tools.
4. Selecting the starting concepts.
5. Selecting and implementing the correct set of methods to translate those concepts into the target language and to construct the synsets and equivalent relations.
6. Selecting and importing the relevant relations into the target wordnet.
7. Constructing a gold standard to which the results can be compared.
8. Use additional methods for evaluation, such as monolingual dictionaries.
9. Devising an evaluation method, such as the confidence score system, applying it and comparing results against the gold standard.
10. Act on the results: Keep methods that are proven to yield good results and/or select those results that reach a certain accuracy, handling the rest manually.
11. Correct all mistakes and fill up all missing links and data for completion of a very small but complete core wordnet.
12. Evaluate this miniature core wordnet, acting on the results where necessary.
13. Extend the wordnet by adding additional starting concepts or another selection, using certain criteria such as adding concepts from the source language that are frequently lexicalised in the target language, or balancing the Top Ontology and/or Domains coverage.

Throughout, processes and results must be documented thoroughly so that intelligent decisions can be made that are based on hard facts. The term “starting concepts” was intentionally used, because a wordnet does not necessarily have to use the Base Concept selection.

Next, we shall investigate each step, starting with the second, in detail and make the relevant recommendations.

4.2.1 *Selecting the source wordnet*

It is generally assumed that, using the expand method, choosing a source wordnet of which the language is conceptually close to the language of the target wordnet benefits the coverage and quality of the latter. Demonstrating this belief, Erjavec and Fišer, in their article, *Building Slovene WordNet*[79], state the following:

> In our case, the notion proposed by Vossen (1998)[80] that a relation holding between two synsets in the Princeton WordNet (PWN) also holds between the corresponding synsets in the new language was taken a step further: we assumed that concepts and relations among them overlap across languages better if the languages are closely related.
> (2006:2)

The author believes that English is conceptually close enough to Afrikaans to warrant being selected as the language of a source wordnet. The reasons are:

- Afrikaans and English are both West Germanic languages.
- English had a substantial influence on the formation of Afrikaans as it is spoken and written today.
- Speakers of Afrikaans and English share the same general geographical space in South Africa (and to a lesser extent, Namibia[81]), and they learn each other's languages, each inevitably having a measure of influence on the other, linguistically as well as culturally. It follows that both languages probably share a great deal of lexicalised concepts.
- Afrikaans has, comparably to English, an advanced vocabulary, used in many different specialised and non-specialised domains.

---

[79] At the time of writing, there was not enough information available on their adopted methodology to warrant investigating it as well in chapter 3. Their article also states in the abstract that their wordnet at that time was still a prototype and contained just about 5000 top-level concepts.

[80] The work referenced is:
Vossen, P. (ed.) 1998. *EuroWordNet: A multilingual database with lexical semantic networks.* Kluwer Academic Press.

[81] Because of political reasons, the number of mother-tongue speakers of Afrikaans in Zimbabwe, which has been small to start with, has declined in recent years. There are also small communities in other African countries such as Lesotho, Swaziland, Mozambique and Botswana, but the numbers are probably too small to take into account here. We shall take as guideline the only country where Afrikaans is an official language, namely South Africa.

Even though the bilingual dictionaries of English and Afrikaans are based on U.K. English, using the PWN should not be a problem because many frequently used British words are included in the latter, and frequently used words from the USA are also included in most South African bilingual dictionaries, often with mention of the British variant or vice versa. Where there could be problems, the author believes that they would be small enough to handle manually.

The PWN 2.1 is available for use. However, according to Dr Ales Horák (personal communication) from the Faculty of Informatics of Masaryk University, Brno, who was involved with developing the wordnet development tools VisDic and DEBVisDic, no wordnets are as yet, at the time of writing, linked to PWN 2.1, and all European languages in the EWN and BalkaNet projects are also still linked to PWN 2.0.

Taking all this into consideration, the conclusion is that the source wordnet will be PWN 2.0.

4.2.2 *Selecting the lexicographic resources and tools*

In chapter 2.3, all possible lexicographic resources at our disposal were investigated. This, of course, does not mean that we are going to use all of them. To refresh the reader's memory, they are:

- *Groot tesourus van Afrikaans* (*GW*) (*Great thesaurus of Afrikaans*)
- *Woordkeusegids: 'n Kerntesourus Van Afrikaans*. (*Word Choice Guide: A Core Thesaurus of Afrikaans*)
- *Handwoordeboek van die Afrikaanse Taal*. (*Desk Dictionary of the Afrikaans Language*)
- *Groot Woordeboek: Afrikaans-Engels, Engels-Afrikaans*. (*Major Dictionary: Afrikaans-English, English-Afrikaans*)
- *Voorsetselwoordeboek Met Engelse Vertalings Asook Enkele Bywoorde*. (*Dictionary of Prepositions with English Translations Including Some Adverbs*)

- List of statistical terms in Afrikaans with English translations, definitions in both languages, as well as contextual information.
- Microsoft Core Terms: A list of computer-related terms in English with Afrikaans translations and English definitions.
- *Brugwoordelys (Engels – Afrikaans) en omskrywings van Afrikaanse begrippe*. (*Word List of Bridge Terms (English – Afrikaans) and descriptions of Afrikaans concepts*)
- (Term list) Bouerswerktuie (*Construction Tools*)
- (Term list) Sterrekundeterme (*Astronomy Terms*)
- *Afrikaanse Woordelys En Spelreëls* (*Afrikaans Word List and Spelling Rules*)
- *Puk/Protea-Korpus Geskrewe Afrikaans* (*Puk/Protea Corpus of Written Afrikaans*)
- *NWU Bible Corpus* (Afrikaans, English and Dutch)

Firstly, one source is crucial to the success of the project: The *GW*, which may be the only bilingual source available in electronic format. It goes without saying that it should be used as the main source for generating cross-linguistic equivalence relations.

For evaluation purposes, we have concluded that all monolingual sources can potentially be used:

- *Woordkeusegids: 'n Kerntesourus Van Afrikaans* (*Word Choice Guide: A Core Thesaurus of Afrikaans*): Lists of synonyms with their parts of speech are provided here, but, as mentioned in section 2.3.2, sense distinctions of homonyms are not made. Still, this can be used to potentially increase confidence scores of already constructed synsets. The exact amount of increase depends on the level of ambiguity. Compare section 3.3.1 for a similar approach in the gloss matching procedure in the context of MultiWordNet.
- *Groot Tesourus van Afrikaans* (*Great Thesaurus of Afrikaans*): This could potentially be used to verify synonymous relations and possibly domain assignments as well. It is, however, not purely synonymously based, as was

mentioned in section 2.3.1. It might be possible to determine, after some preprocessing, which groups of words are synonymous and which are not, but this might involve some complicated procedures. However, if words from the same Afrikaans synset occur here in the same list, its confidence score can be raised slightly because of the relative chance of any two words in a list being synonyms. The bigger the list, the smaller the increase.

- *Handwoordeboek van die Afrikaanse Taal* (*Desk Dictionary of the Afrikaans Language*): This is a rich resource for various kinds of evaluation: synonyms, variants, related words, domains, hypernyms and gloss matching.
- *Afrikaanse Woordelys En Spelreëls* (*Afrikaans Word List and Spelling Rules*): This is the official source for checking the spelling of literals. If any word in an Afrikaans synset does not occur here, it must be flagged for inspection.
- *Puk/Protea-Korpus Geskrewe Afrikaans* (*Puk/Protea Corpus of Written Afrikaans*): This can be used to support the addition of words to the Afrikaans wordnet, mapping them to PWN synsets, or the exclusion of words from already existing Afrikaans synsets, by providing frequency information. The validity of word choices and collocations in glosses can be verified through the corpus as well.
- *NWU Bible Corpus*: The translation pairs with their probability scores can be used for evaluation, but only if they match with a translation pair in the bilingual dictionary and only if the Afrikaans word occurs frequently in the *Puk/Protea Corpus*. In those cases, a match can raise the confidence score of a synonymous assignment.
- (Term lists) Terms are by definition used in specific domains only, although some of them might be general enough to be included in a wordnet. If a word is not found in the *AWS*, these lists might be consulted in conjunction with the corpus to possibly determine their correct spellings and to suggest inclusion or exclusion by way of their frequency counts.
- *Voorsetselwoordeboek Met Engelse Vertalings Asook Enkele Bywoorde* (*Dictionary of Prepositions with English Translations Including Some Adverbs*). This can be used to extract typical collocations and their translations, which is useful especially if translation equivalents cannot be found for any multiword phrases present in PWN synsets.

In section 2.4.1, the tool DEBVisDic is described. The author is not aware of any other quality wordnet development tools that are freely available, except for its predecessor VisDic, which was used with great success in the BalkaNet project. Both VisDic and DEBVisDic greatly facilitates many tasks, although they do not possess certain functionalities, such as automatically selecting and exporting a group of synsets based on certain criteria (structural, content, etc.), calculating conceptual distance, etc. Some of these can be done by writing queries in, for example, SQL; otherwise, programmes must be written to handle these procedures, which can, for example, be done in Perl.

Many tasks consist mainly of processing, sorting and comparing data, which can be done by relatively simple programmes and queries. The rest mainly involve constructing synsets, importing or creating links, or importing, exporting or converting files. The latter group is well covered by the functionalities of VisDic and DEBVisDic. It is therefore of the author's opinion that other tools are currently not needed.

4.2.3 *Selecting the starting concepts*

To select a list of starting concepts for building a wordnet, possible approaches are to select a set of the most frequent words in the language using corpora or using a set of lemmas found in a standard monolingual dictionary. This is perfectly feasible; however, these words then have to be mapped to their counterparts in the source wordnet, if the expand methodology is to be used.

Fortunately, a direct selection of possible starting concepts from the PWN has already been made, in the form of the so-called Base Concepts (BCs) introduced in the EuroWordNet project (see section 1.2.1). Their validity as general, frequent concepts have been proved to a degree by the application of the criterion that they must occur in at least two of the languages involved, as well as being high up in the semantic hierarchy and having relations to many other concepts (http://www.globalwordnet.org/gwa/gwa_base_concepts.htm).

The BalkaNet project has extended the 1024 Common BCs[82] (CBCs) to 4689 and upgraded the mapping of the EWN CBCs, which were defined as PWN 1.5 synsets, to PWN 2.0. These 4689 synsets are downloadable from http://www.globalwordnet.org/gwa/gwa_base_concepts.htm.

Prof Vossen has suggested (personal communication) that building a small core wordnet of about 5000 synsets can take about a year to complete. Although the project team should be able to work on the wordnet for at least this period of time, some kind of result must be shown after three months. It is therefore suggested that a portion of the available Base Concepts be selected in order for the team to be able to complete a full cycle in the construction process for production of a very small core wordnet. This portion comprises the so-called BalkaNet Concept Set 1 (BCS1), which, together with the bigger BCS2, make up the selection of the downloadable 4689 Base Concepts. BCS1 consists of 1218 synsets, most of which are Common BCs from EWN and are therefore almost guaranteed to represent a section of the Afrikaans vocabulary that is very generic and frequently lexicalised.

Furthermore, it is suggested that at least the first selection consists of nouns only, since many methods that were discussed were only used on words from this category. Therefore, the very first set of starting concepts is the nouns from BCS1.

4.2.4 *Selecting and implementing the correct set of methods to translate concepts into Afrikaans and to construct the synsets and equivalent relations*

4.2.4.1 *Selection of the methods*

In chapter 3, several methods were introduced to translate source wordnet concepts into the target wordnet language and using these translation equivalents, possibly with additional resources, to construct the synsets in the target wordnet, each one being aligned via equivalence relations to at least one synset in the source wordnet language. The following table summarises the presented methods or heuristics (those of the Dutch WordNet excluded), where SW denotes “source wordnet”, TW denotes

[82] Common Base Concepts act as BCs in at least two languages, as opposed to Global Base Concepts acting as BCs in all languages of the world, and Local Base Concepts that act as BCs in one language only (http://www.globalwordnet.org/gwa/gwa_base_concepts.htm).

"target wordnet", TE "translation equivalent", SL "source language" and TL "target language":

Wordnet	**Method/Heuristics**	**Description and possible example**
Spanish	class (1)	"use as knowledge sources individual entries coming from bilingual dictionaries and WN synsets" (Atserias et al 1997:2): (1) two or more variants in SW synset with a single translation in TL means the latter is linked to the SW synset
Spanish	class (2)	"use as knowledge sources individual entries coming from bilingual dictionaries and WN synsets" (Atserias et al 1997:2): (2) If two or more SL words with the same field identifiers occur in the same synset, they are linked to all of their Spanish translations.
Spanish	structural (1)	"take profit of WN structure" (Atserias et al 1997:2): (1) If all TEs of a TL word share at least one common synset in

		its wordnet, link TL word to all common synsets of its translations.
Spanish	structural (2)	"take profit of WN structure" (Atserias et al 1997:2): (2) If a TE is part of a synset which is a direct parent of the synsets corresponding to the rest of the TEs, the TL word is linked to all hyponym synsets.
Spanish	structural (3)	"take" profit of WN structure (Atserias et al 1997:2): (3) If all TEs have synsets which are brothers respecting to a common parent, the TL word is linked to all co-hyponym synsets.
Spanish	structural (4)	"take profit of WN structure" (Atserias et al 1997:2): (4) If the synset of a TE is a distant hypernym of synsets of the rest of the TEs, the TL word is linked to the lower-level (hyponym) synsets.
Spanish	Conceptual Distance (CD)	"makes use of knowledge relative to meaning closeness

		between lexical concepts" (Atserias et al 1997:2): e.g. computing the CD between two assigned synonyms to determine the validity of this assignment
Italian (MultiWordNet)	generic probability	calculates confidence score (CS) depending on number of candidate synsets for a given TL word or sense
Italian (MultiWordNet)	back translation	raises CS if a synonym of a TE of a TL word can be translated back to that TL word
Italian (MultiWordNet)	gloss matching	matches gloss information of TL dictionary entries with SW counterparts; if matches found, raises CS accordingly
Italian (MultiWordNet)	synset intersection	increases CS for SW synsets containing more TEs of a TL word than other synsets
Romanian (Barbu and Mititelu)	first heuristic rule (class/structural)	If one or more words in a SW synset are monosemous, its TW equivalent consists of all TEs of the monosemous word or words, otherwise it consists of

		all TEs that are linked to all the words in the SW synset.
Romanian (Barbu and Mititelu)	second heuristic rule	determines if two different source synsets that stand in a hypernym/hyponym relation to each other can be mapped to a single target synset
Romanian (Barbu and Mititelu)	third heuristic rule	WordNet Domains are used here to help construct TL synsets and map them to SL synsets. A lot of preprocessing was needed.
Romanian (Barbu and Mititelu)	fourth heuristic rule	TL synsets are generated based on criteria of similarity between the glosses of their individual words with those of the words in the SL synsets. Preprocessing was required in the form of automatically lemmatising and tagging all relevant glosses and constructing vectors from their nouns, followed by some more computational

		processing to be able to finally generate a synset.

Figure 9: Table summarising wordnet construction methods and heuristics presented in chapter 3

Considering that there is a strict time limit, as well as the abundance of automatic methods available, some of the above heuristics or methods can be discarded right away:

- Italian wordnet (MWN), gloss matching: Complex processing of dictionary glosses is required, which is time-consuming.
- Romanian wordnet, heuristics 3 and 4: A lot of preprocessing is needed here, as well as relying on a high quality dictionary, which is not proven without a doubt in our case.

As for the rest, notice that in the Spanish WordNet, in all cases, except for Conceptual Distance and class (2) (where the latter involves more than one TL word), a link is immediately made between a TL word and one or more SW synsets, whereas in MWN, confidence scores (CS) are given to SW synsets that can be linked to one or more TL words. The Romanian wordnet heuristics are similar to the Spanish methods in the sense that translations and links are directly made, but they differ in the fact that synsets are always constructed in one step.

Considering the Spanish methods, some of them can also be eliminated:

- class (1): Notice the similarity here with Barbu and Mititelu's first heuristic rule, except that the latter achieves more (the synset is constructed in one step) and is logically more intuitive (it is highly likely that the synset of a monosemous word, irrespective of its synonyms, must have as equivalent in the TW the set of TEs of the monosemous word). Note that, obviously, variants must be considered as sharing the same set of TEs.

- class (2): The author believes that two SL words in a synset sharing the same field identifier do not increase the precision of the set of possible TEs; at least it is not intuitive. Of course, the possibility remains that the dictionary field identifiers of the literals in already constructed synsets can be compared to the WordNet Domains in their respective linked synsets, resulting in a confidence score. This, however, depends on the available time and quality of the dictionary or dictionaries.
- structural (1): This is also very similar to the second part of Barbu and Mititelu's first heuristic rule, where a TL synset consists of all the TEs that can be translated to all the literals in the SL synset. The latter is, however, dependent on a condition, namely that no word in the SL synset must be monosemous, and is therefore more likely to be accurate.

Regarding "structural (2)" and "structural (4)", this is reminiscent of the scenario described in Barbu and Mititelu's second heuristic rule, where three SL synsets in a hypernym/hyponym relation are mapped to two TL synsets in the same relation, resulting in choosing the lower TL synset. It is obvious in all three cases that more a specific (more information) rather than more general (less information) equivalence is preferred. Moreover, the Spanish methods seem to get it right the first time, where their equivalent in the Romanian wordnet is a rectification of an already constructed equivalence relation. Because the conditions are different in all cases, these three methods are still recommended.

The third structural method is recommended as well, because of its success and because the specific scenario (of co-hyponyms) is not represented elsewhere.

Conceptual Distance (CD) will be used to verify already established equivalence relations. Because it calculates paths in the hierarchies of both wordnets, it can only be used when the Afrikaans wordnet is in an already more advanced stage. Based on figure 4 in section 3.1.1 on using CD in construction of the Dutch WordNet, the conditions, assuming that nouns are the only word category in question, for determining the CD between the sense of a word and all its equivalents in the other wordnet are:

- Both the SW synset in which the word sense appears and the synsets of all its possible equivalents must be extended to at least one lower level in the hierarchy, i.e. all of their hyponyms must be represented.
- Both the SW synset in which the word sense appears and the synsets of all its possible equivalents must have hypernyms up to the top of the hierarchy.
- Both the SW synset in which the word sense appears, as well as all of its possible direct hypernyms and hyponyms must already have equivalence relations established.

If all of the above criteria are met, then it would be possible to determine which translation equivalents of a given word sense are conceptually closer to it than others. The word sense that is closest should be in the synset that is linked as equivalent to the SW synset.

It can also be used to determine if TEs in the same TL synset have the right senses assigned to them. This was explained in section 3.1.1 where the Dutch word *orgel* can be translated to both *organ* and *keyboard*. Using this example, the CD between these different senses of *organ* and the particular sense of *keyboard* is calculated, so that the sense that is closest can be chosen or verified as a suitable synonym of *keyboard*. Note that in this case, the particular senses of *orgel* and *keyboard* must already be verified as being equivalent.

Regarding the MWN methods, the methods of generic probability, back translation and synset intersection are logically based and can be achieved fully automatically. Hence we also recommend these. It follows that a confidence score system must also be devised for this purpose.

The first and second heuristic rules of Barbu and Mititelu can also be applied automatically. However, since they already produce synsets, the results must be compared against a gold standard instead of using a confidence score.

The author believes that the MWN methods can be successfully combined with the Spanish and Romanian methods and that they can be made to complement each other. This is discussed in the next section.

4.2.4.2 *Implementing the methods*

First of all, we consider merging both directions of the bilingual dictionary to create a "homogeneous bilingual (HBil)" as was mentioned in Atserias et al (1997:2). This entails getting all TEs in one direction and then obtaining all the TEs of the first set of TEs. However, since TEs are not assigned sense numbers, they will have to be assigned automatically when synsets are built, using an *Hbil*. Having all literals assigned senses by the dictionary beforehand greatly simplifies using the MWN procedures, because it reduces the set of TEs in each case and because in some cases one does not have to account for ambiguity; therefore, it is suggested that we do not create an *HBil*. Using the MWN procedures first, this will eventually result in every Afrikaans word sense having a confidence score in relation with one or more PWN synsets. Conversely, every involved PWN synset will have a list of TEs, each of which is assigned a CS in relation to the synset.

4.2.4.2.1 *Applying the MultiWordNet methods and the confidence score system*

Because the MWN methods link Italian senses to the PWN and not the other way around, we first have to obtain a list of all possible TEs from all the starting concepts. The ratio between those literals that could be translated automatically and those that could not is referred to as the *recall*. Usually, in conjunction with this, the term *precision* is used. In this case, it would denote the ratio between correctly and wrongly translated literals. However, since the quality of the bilingual dictionary is not under suspicion, and since it is, for the time being, the only bilingual resource electronically available, we shall assume that all successfully translated literals are correct. For those that could not be translated, the following steps are followed, which is loosely adapted from a procedure described in the article *Coping with lexical gaps when building aligned multilingual wordnets* by Bentivogli et al (2000):

1. Automatically search for translation equivalents under the usage section, where typical collocations, idioms and other typical phrases are displayed.

2. If this is not successful, consult the *Voorsetselwoordeboek* (*Dictionary of Prepositions*) for possible automatic translation.
3. If there remain untranslated literals, translate these manually. For genuine lexical gaps, the PWN synset may be left untranslated.

The above work of Bentivogli et al aims at reducing the "manual work necessary to cope with lexical gaps in the construction of aligned multilingual wordnets" (2000:1) and contains more steps than listed here, but some of them involve more complicated methods such as the semi-automatic recognition of certain phrasal groups in the dictionary. We are not going to consider these methods because of the lack of time and resources.

After consulting the *AWS* to check the spelling of all generated Afrikaans words, the next step entails obtaining a set of PWN candidate synsets for every translation equivalent. These synsets may not all belong to the original set of Base Concepts, since some of the words may have additional senses and/or may be back-translated differently. All synsets in which one or more TEs of an Afrikaans word sense occur are considered candidate synsets.

After obtaining the above set, one can apply the chosen methods. For each one, a confidence score is calculated, after which some kind of average is produced per PWN synset in relation to an Afrikaans word. Compared to the other accepted MWN methods, the generic probability method is less important and should therefore carry less weight in the final CS calculation. The reason is that the number of synsets in the candidate set should influence the choice of a certain synset less than the number of translation equivalents it contains, for example. It is therefore suggested that this method carries a 10% weight, while the other two accepted MWN methods each carry a 45% weight.

Implementing the generic probability method, the following steps are followed to calculate the confidence score:

1. Assume the maximum CS is 95%, for only one candidate synset. This is because an equivalence assignment can never be considered correct without a

doubt if it is not confirmed manually, even if the SW synset is the only candidate. For each Afrikaans word sense, divide 95% by the number of candidate PWN synsets.

2. The CS for this method is assigned the above amount.
3. The formula is: CS = $\frac{95}{CandSet}$ where *CandSet* is the number of candidate synsets.

For example, an Afrikaans word sense may have three candidate synsets. Therefore, the CS for each synset is 95/3 = 31,67%.

The back translation method calculates a CS depending on the number of literals in a candidate synset that can be translated back to the Afrikaans word in question. The total number of literals in the candidate synset must also be considered. Once again, the maximum possible CS is 95%. This is for the case of all elements in the synset being back-translatable to the Afrikaans word. Also, because at least one word in the synset can be back-translated (since it is already a candidate synset) there already exists an initial CS. Although strictly speaking, the chance that an English word can be translated back to a different sense increases with a bigger number of Afrikaans senses, this factor should have minimum impact because the right set of TEs was already chosen for the Afrikaans sense via the bilingual dictionary. Also, it should be highly unlikely that a synset has a high CS for a wrong Afrikaans word sense using this method. The following algorithm is implemented:

1. Divide 95% by the number of literals in the PWN synset.
2. Multiply this by the number of literals that can be translated back to the Afrikaans word.
3. The CS is assigned the above amount.
4. The formula is: CS = $\frac{95t}{l}$ where *l* is the number of literals in the synset and *t* the number of literals that has the Afrikaans word as one of their TEs.

Calculating the CS for the synset intersection method is also straightforward. For a given Afrikaans word sense, the CS increases if more synonyms in a candidate synset match with TEs of the Afrikaans word in the dictionary. The algorithm is:

1. Divide 95% by the number of literals in the PWN synset.
2. Multiply this by the number of literals in the synset that match with the translation group in the dictionary for the Afrikaans word sense in question.
3. The CS is assigned the above amount.
4. The formula is: CS = $\frac{95t}{l}$ where l is the number of literals in the PWN synset and t the number of literals that match the translation group of the Afrikaans word sense in the dictionary.

After the CS for all three methods are calculated, the final CS is obtained by the following formula:

Final CS = $\frac{1 \times CS1}{10} + \frac{9 \times CS2}{20} + \frac{9 \times CS3}{20}$ where $CS1$, $CS2$ and $CS3$ are the confidence scores obtained by the three methods respectively.

This formula must be applied to every synset in relation with every Afrikaans word sense. For example, if synset A obtained a CS of 33% for linking with Afrikaans sense S using generic probability, and the same synset obtained a CS of 75% and 90% respectively for the same Afrikaans sense S using the other two methods, the formula is used to derive the final CS for this pair, yielding 3,3% + 33,75% + 40,5% = 77,55%.

4.2.4.2.2 *Adding the methods and heuristics from the Spanish WordNet and the Romanian WordNet*

After the MWN methods are applied, each PWN synset has a CS in relation with a certain Afrikaans word sense. The idea is to apply the other methods such that both the separate confidence scores and the gold standard can be used to verify synset equivalent relations and the choice of literals. One potential problem is that the Spanish and Romanian methods are not applied to TL senses, though, but TL words.

However, the Spanish methods can easily be used on senses as well. Also, each method can potentially override an equivalence relation that another has made, therefore care must be taken not to favour one method that has not been proved as valid at the cost of another one.

Our answer to this is to apply all these methods in parallel and finally to compare all the results to a gold standard. This is done in the following way:

1. First, get four identical sets of both the Afrikaans words and the PWN synsets in which their TEs occur.
2. Using the first set, obtain all confidence scores using the MWN methods as described in the previous section. The deliverable of this step is a list of all the relevant PWN synsets, each with a list of Afrikaans word senses in decreasing order of CS.
3. Then, using the second set, obtain the TEs of their senses instead of just the words, and apply the Spanish methods, *structural* 2-4, in that order. The deliverable of this step is a list of all relevant PWN synsets, each with a set of Afrikaans word senses linked to it. The latter can be regarded as preliminary Afrikaans synsets.
4. Using the third set, apply Barbu and Mititelu's heuristic rule number one to the list of PWN synsets. Afrikaans word senses cannot be used here, since a PWN synset is used as the basis to extract all possible Afrikaans equivalents. The deliverable of this step is a list of all relevant PWN synsets, each with a set of Afrikaans words linked to it. The latter can also be regarded as preliminary Afrikaans synsets.
5. Using the fourth set, apply Barbu and Mititelu's heuristic rule number two to the list of PWN synsets. The deliverable of this step is the same as above.

The end result is four lists of PWN synsets, each one linked to one or more Afrikaans words, the first list with an added CS and the first two lists linked to Afrikaans senses instead of just words.

4.2.5 *Constructing and applying the gold standard*

Instead of comparing the synsets from the different lists to each other, they are all compared to a gold standard, eliminating the danger that a bad method be used. The gold standard must first be constructed. This is done in the following way:

1. Choose 200 noun synsets from BCS1. They must all be linked to each other via semantic relations, and all their hypernyms to the top of the hierarchy must be included.
2. Construct their Afrikaans equivalents manually.
3. Each one must be verified by a qualified lexicographer.
4. Use these synsets and their Afrikaans TEs in the above methods and heuristics.
5. Compare the last three deliverables to the gold standard, using the scenarios given in Barbu and Mititelu (2005:102). Sense numbers are ignored here. They are:
    a. The synsets are equal (Identical).
    b. The generated synset has all literals of the correct synset and some more (Over-generation).
    c. The generated synset and the golden one have some literals in common and some different (Overlap).
    d. The generated synset literals form a proper subset of the golden synset (Under-generation).
    e. The generated synset have no literals in common with the correct one (Disjoint).
6. As in the abovementioned work, the cases of Over-generation, Overlap and Disjoint are counted as errors.
7. A synset derived from an MWN method is regarded as an error if:
    a. A literal from its equivalent in the gold standard is missing from the list of word senses with confidence scores
    b. The literals in the gold standard synset have average confidence scores in the list of less than 80%.
8. If the number of erroneous synsets from a given method is too high (low precision), consider discarding the method. This is done if no means can be found to improve the method.

Step number 7 and 8 here are especially crucial to the success of the automatic procedures used in building Afrikaans synsets and equivalence relations. This also applies to the confidence scores given to synsets and word senses using the MWN methods: If, for a given Afrikaans synset in the gold standard, its literals do not have relatively high confidence scores, and this occurs often enough, this could mean that the CS system is not accurate enough. If, however, the confidence scores of the literals in the gold standard synsets are generally good, this further strengthens the suggestion that the particular synset and equivalence relation is correct. However, the CS system is not strictly necessary, since a gold standard, which is more accurate, is constructed. It is suggested that an average CS of 80% or more in a synset is acceptable, while erroneous assignments amounting to more than 20% of all selected synsets warrants investigation into the methods used. Such an investigation will imply looking at single methods, such as back translation, in isolation and comparing only their results against the gold standard.

If the above process of eliminating bad methods is complete, the following must be done:

1. If the MWN methods are successful, continue using them for the rest of the Afrikaans word senses.
2. Build synsets from all other methods that passed the previous test. If two or more different Afrikaans synsets are produced by different methods from the same PWN synset, select the result of the method that scored the highest in the gold standard, but only if its literals from the MWN methods have average confidence scores exceeding 80%. If this is not the case, these steps are followed:
    a. Select the other method, but also only if the confidence scores have an average of more then 80%, and if the method itself had less than 15% errors when compared against the gold standard.
    b. If this is not the case, flag the PWN synset for manual translation.
3. Automatically build word senses according to their occurrence in the Afrikaans synsets. For example, the first time the word *groot* (great, big, large) is put in an Afrikaans synset, assign it the sense number 1. The second time it

is assigned the number 2, etc. The numbers can differ from their sense numbers in the dictionary because they can reflect, at least originally, a finer or coarser sense granularity in the PWN, since this is a relatively strict expand methodology.

Depending on the success of the previous procedures, Afrikaans synsets can further be evaluated using the thesauri. If the confidence score system is flawed, another one can be devised, if necessary, using synonym sets from these resources. This possibility was not included in the first phase because of the relative complexity involved. Another possibility is constructing a bigger gold standard, increasing the likelihood of the success of automatic procedures after evaluation.

4.2.6 *Importing the semantic relations*

If the synset construction and equivalence linking algorithms are finally tuned in, the semantic relations can be imported. For this purpose, we use the algorithm provided by Barbu and Mititelu (2005:105) that was described in section 3.4.6. The completion of this will result in a miniature core wordnet of nouns containing only semantic relations.

4.2.7 *Evaluating the first version of the core wordnet*

If the core wordnet is small enough, it can be evaluated manually; otherwise, a set of around 200 interconnected Afrikaans synsets is selected. The equivalence relations and literal selection are evaluated once again, as well as the newly established semantic relations. Regarding the latter, the following changes can be made:

- Adding another synset: This is reserved for making a distinction between concepts where it is not made in PWN. The new synset and its closest equivalent, from which it is derived, are usually both linked to the same PWN synset.
- Deleting a synset: This is done when it is found that inclusion of its contents is not warranted.
- Merging two or more synsets: When a sense distinction is made in PWN that is not done in Afrikaans, this can be done, resulting in a new synset linked to

two or more PWN synsets. This scenario is covered by the second heuristic rule of Barbu and Mititelu, but its successful application is, of course, not guaranteed.

- Changing a synset: Some literals in a synset can be changed when it is found that their selection is not warranted. For example, a reason could be that they are not included in a monolingual dictionary.

In all of the above cases, changes in sense numbers must be considered and be made to reflect accordingly.

After the manual evaluation, the Conceptual Distance method can be used to verify cross-lingual equivalence relations, as well as Afrikaans synonymous relations within synsets.

4.2.8 *Adding lexical relations*

This is a completely different step, which may involve different procedures. Firstly, a study must be made to determine which relations can be imported automatically. For this purpose, one could make another small selection of PWN concepts whose literals are linked via lexical relations. If the automatic import of a certain type of lexical relation is not possible, these steps are followed, where we assume that nouns, verbs, adjectives and adverbs already exist:

1. For all antonyms (nouns, verbs, adjectives and adverbs), search for monolingual gloss information that specify this. For example, under the lemma “groot” (big), the gloss can contain the phrase “teenoor *klein*” (as opposed to *small*), revealing an antonymic relation. The optional tag “sien” (see also) could also indicate an antonym.
2. The links *derived from*, *relational adjective* and *participle* may be derived automatically through regular morphological variation, examples of which may be found within the entry itself, such as *music* and *musical*.

The lexical relations that cannot be derived automatically must be dealt with manually.

4.2.9 *Extending the wordnet*

The first miniature core wordnet can be extended in two possible ways:

- Extend the PWN noun selection, using the methods that were proved as valid.
- Make another selection of words from another word category for expansion to the Afrikaans wordnet, such as an additional set of verbs from the BCS1.

It would certainly be easier to choose the first option, as one need just continue in the same fashion as before. For another word category, the selected methods must be evaluated again and a new gold standard must be devised. It is possible that additional methods for translation and synset construction must be used. However, by adding another category, one could claim, by completion of the first selection, that the wordnet is more complete.

The author believes that, rather, the set of nouns must first be extended to at least a couple of thousand. Other categories may refer via lexical relations back to nouns, and if these nouns do not exist yet, it may complicate the process of adding them later on.

Building an extensive wordnet of nouns is also a practise that has been followed by others. In the on-line article, *Portuguese WordNet general architecture and internal semantic relations* (Maraffa 2002), the author states that the Portuguese WordNet at that time consisted of approximately 10 000 distinct noun forms and also implicates, in the following sentence, that building a wordnet one category at a time is less time-consuming:

> Due to operative and resource economy reasons, on the one hand, and to applications goals, on the other hand, the first version of the Portuguese WordNet is mostly focused on nouns.
> (Maraffa 2002)

The methods presented by Atserias et al (1997) are also just valid for nouns. So are the heuristic rules devised by Barbu and Mititelu for their version of the Romanian WordNet.

Taking all this into account, it is suggested that the noun selection be extended to the BCS2 and BCS3. Afterwards, another word category may be chosen, or the coverage of the Domains, the SUMO[83] or EWN Top Ontology, when applied, could be balanced out with more nouns to attempt as wide a coverage as possible by the core wordnet. This could also be achieved by selecting a set of words that are lexicalised frequently in Afrikaans but are not yet included.

4.2.10 *A preliminary work schedule*

Besides other duties such as training employees to use tools such as DEBVisDic, the general work schedule may look as follows:

1. The lexicographers obtain all relevant lists of words and synsets in both languages for the first cycle, after which they start to construct the gold standard.
2. At the same time, the computational linguists write the programmes necessary to achieve tasks such as applying synset intersections, calculating confidence scores and conceptual distances, searching the bilingual dictionary for specific strings, etc.
3. If some Afrikaans words are not found in the *AWS*, the lexicographers must check other sources and decide whether to include them or not.
4. The lexicographers must also handle any translation problems, such as lexical gaps, using the algorithm described in section 4.2.4.2.1.
5. Apply the chosen set of methods in the described order using the programmes and DEBVisDic. Any manual work must be done, as far as possible, by using the latter.
6. Obtain results in table format of success rates, such as precision, recall and confidence scores, with mention of the methods used.

[83] SUMO = Suggested Upper Merged Ontology. This is an ontology that is integrated in PWN 2.0 and which contains elements such as Action and Process that can potentially be used for word sense disambiguation and categorisation.

7. Evaluate the results and decide on a suitable action, guided by the principles stated in section 4.2.5.
8. Investigate additional methods of evaluation, if time allows, by using monolingual dictionaries or thesauri, for example.
9. Keep to a time schedule: Plan to construct and/or evaluate a certain number of synsets and equivalent relations per day, depending on the deadline. Try to establish the procedures within a certain time frame, for example in one month. The time limit may influence the selection of methods to be used.
10. Keep in contact with the collaborators, heeding their advice and exchanging useful information, as well as being explicit about everyone's role in the project.

The second last step is a reminder that it is important to be flexible. The methodology and work schedule presented here must be regarded as a preliminary guide only. If more accurate methods can be devised, it must be investigated and applied, if time allows.

Chapter 5: Final discussion

5.1 Expected results and problems

It is always good to remain optimistic and expect good results. But one also needs to be realistic and try to anticipate any future problems. First of all, we look at the results that can be expected after the first phase, which is the period before the first deadline, i.e. three months:

- We expect to have built a small core wordnet of nouns, consisting of at least 2000 synsets.
- The methodology and selection of resources must have been refined and standardised by then.
- We must have investigated methods that can be applied to other word classes by then.
- We must have a clearer vision of future extension and eventual application of the wordnet.
- All documentation must be updated and, preferably, at least one article must be written on the project.

Some foreseeable problems are:

- A possible lack of quality of the lexicographic resources, resulting in more manual work.
- There may be more conceptual differences between English and Afrikaans than anticipated.
- The procedures presented in chapter 4 may not produce the desired results, potentially resulting in a loss of accuracy and/or more manual work.
- The methodology may prove too complicated, meaning that it must be simplified, which can also result in a loss of accuracy and/or more manual work.

Only time will tell whether the problems will materialise and whether the expected results will come to fruition.

5.2 Results of research

The research has not only revealed some very interesting facts about the world of wordnets, but has also shown, to some degree, that Afrikaans can be represented in this world and how it could possibly be done. It has also made clear that although many more automatic procedures exist for constructing wordnets than around a decade ago, much still depends heavily on quality resources and tools, as well as idiosyncrasies of the languages involved. The author believes that there could still be room for improvement and that many resource-scarce languages will take profit of more innovative ways to make use of their resources in constructing wordnets.

5.3 Conclusion and suggestions for future work

This work is a humble attempt to propose a methodology for building a wordnet for Afrikaans. It suggests combining tried and tested methods from different wordnets and extracting the best results for maximising high quality automatic processing as far as possible. Quality evaluation procedures are also proposed, keeping in mind the limited lexical resources, time frame and budget. It is hoped that by this work, at least one small step can be taken in the direction of making wordnet construction for resource-scarce languages easier, and that as a result, the state of natural language processing (NLP) in South Africa can be improved.

Some obvious suggestions for future work entail possible applications of the wordnet. Here are just a few possibilities:

- Using the wordnet to build a semantically tagged corpus, the two resources could be combined to be used in word sense disambiguation, machine translation, document classification and other NLP tasks.
- The wordnet can be used for information retrieval on the Internet and other database systems.
- At the moment we are working on a morpho-syntactic lexical database for Afrikaans, which will be linked to the wordnet. The two resources combined could lead to increased accuracy for NLP tasks, as well as enabling the

development of better quality core technologies such as intelligent spelling, grammar and style checkers.

- The wordnet could be linked to CALL (Computer Assisted Language Learning) software to aid in teaching Afrikaans to speakers of other languages. This will become even more effective as the coverage is increased and the database is extended to other South African languages. The latter is an important goal regarding the high priority of language development in South Africa.
- A possible long-term goal is the development of a knowledge base, which, in conjunction with the wordnet, will be of great use in various NLP tasks.

Obviously, the wordnet can be improved and extended in the future. Besides the obvious plan to extend the coverage, some other possibilities are:

- Future linking to an ILI to be part of a possible global wordnet database.
- Extension to the other South African languages using a database for multilingual wordnets.
- According to a presentation by Christiane Fellbaum and George A. Miller[84], the Princeton University is planning to upgrade the PWN to a so-called WordNet Plus that will have the following features:
    - Connecting all synsets to each other via so-called "directed, weighted arcs", which represent association or evocation between concepts (e.g. dollar -> green).
    - More semantic relations (e.g. "axe – tree", "buy - shop")
    - More links across parts of speech (e.g. "traffic, congested, stop")
    - The possible introduction of an "annotator robot" which can learn to rate evocations like a human.

  If the WordNet Plus model can be extended to other languages, this can be applied to a possible future Global WordNet. At this time it will be even more advantageous for the Afrikaans wordnet to upgrade to obtain these features,

[84] *Whither WordNet?* http://www.lrec-conf.org/lrec2006/IMG/pdf/AZPrize.Christiane Fellbaum Presentation.LREC06.pdf

not only for improved monolingual NLP performance, but also for better cross-lingual processing and research.

5.4 Final word

The wordnet phenomenon is one of the highlights of the human language technology scene today. It is with great wonder that one can look back at the almost exponential development of language technologies in the last couple of decades. The worldwide collaboration and exchange of information regarding wordnet and wordnet related matters that are prevalent today were certainly unprecedented and unexpected at the time when the first version of the Princeton WordNet was released. The fact that also "smaller" languages with limited lexical resources do not have to stand in the back row regarding NLP is a positive sign for the future development of these languages and the increasing role that they can play in international research and technology. We hope that the Afrikaans wordnet will be part of a great series of advances that can make these dreams come true.

BIBLIOGRAPHY

Atserias, J., S. Climent, X. Farreres, G. Rigau, and H. Rodriguez, 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In: *Proceedings of "Recent Advances on Natural Language Processing" RANLP'97*, Tzigov Chark, Bulgaria. http://cv.uoc.es/~grc0_001091_web/files/atserias.pdf

Barbu, Eduard and Mititelu, Verginica Barbu. Automatic Building of Wordnets. In: *Proceedings: International Conference – Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgaria. 21-23 September 2005. (https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2005/RANLP/papers/89_barbu.pdf)

Beckwith, Richard and Miller, George. Implementing a Lexical Network. In: *International Journal of Lexicography* 3(4), 1990, pp. 302-312.

Bentivogli, L., Pianta, E. and Pianesi, F. (2000) *Coping with lexical gaps when building aligned multilingual wordnets.* Proceedings of LREC 2000, Athens, Greece, pp. 993-997. http://www.multiwordnet.itc.it/paper/wordnet-lrec2000.pdf

Clark, P., Harrison, P., Jenkins, T., Thompson, J. and Wojcik, R. From WordNet to a Knowlege Base. In: *Proc AAAI 2006 Spring Symposium on Formalizing and Compiling Background Knowledge*, 2006. http://www.cs.utexas.edu/users/pclark/papers/ss06.pdf

Daille, Béatrice. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act Combining Symbolic and Statistical Approaches to Language, MIT Press (1995). http://acl.ldc.upenn.edu/W/W94/W94-0104.pdf

Erjavec, Tomaž and Fišer, Darja. Building Slovene WordNet. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06*. 24-26 May 2006. Genoa, Italy. http://nl.ijs.si/slownet/bib/slown-LREC05.pdf

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press.

Horák, A., Pala, K., Rambousek, A. and Povolný, M. DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In: *Proceedings of the Second International WordNet Conference – GWC 2004,* pages 136-141, Brno, Czech Republic, 2003.

http://nlp.fi.muni.cz/publications/gwc2006_hales_pala_etal/gwc2006_hales_pala_etal.pdf

Lenat, D., Miller, G. and Yokoi, T. CYC, WordNet and EDR: Critiques and Responses. In: *Communications of the ACM*, Volume 38, Issue 11 (November 1995). pp. 45-48.

Marrafa, Palmira. 2002. *Portuguese WordNet general architecture and internal semantic relations*. São Paulo. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502002000300008&lng=es&nrm=iso&tlng=en

Morato, J., Marzal, M., Lloréns, J. and Moreiro, J. WordNet Applications. In: *GWC 2004, Proceedings*, pp. 270-278. Sojka, Petr, Pala, Karel, Smrž, Pavel, Fellbaum, Christiane, Vossen, Piek (eds.). ©Masaryk University, Brno, 2003. www.fi.muni.cz/gwc2004/proc/105.pdf

Paducheva, Elena V., Rakhilina, Ekaterina V. and Filipenko, Marina V. 1992. Semantic dictionary viewed as a lexical database. In: *Proceedings Of COLING-92, Nantes, Aug. 23-28, 1992.* http://www.lexicograph.ru/eng/pub/files/16-nant.pdf

Pianta, E., Bentivogli, L. and Girardi, C. MultiWordNet: Developing an aligned multilingual database. In: *Proceedings of the 1st International WordNet Conference, January 21-25, 2002, Mysore, India, pp. 293-302*. http://multiwordnet.itc.it/paper/MWN-India-published.pdf

Sowa, John F. 1992. *Semantic Networks*. http://www.jfsowa.com/pubs/semnet.htm

Tufiş, D., Cristea, D. and Stamou, S. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In: *Romanian Journal of Information Science and Technology*. Volume 7, Numbers 1-2, 2004, 9-43. www.ceid.upatras.gr/Balkanet/journal/7_Overview.pdf

Vossen, P., Bloksma, L., Boersma, P. 1999. *The Dutch WordNet*. http://www.vossen.info/docs/1999/DutchWordNet.pdf

Vossen, Piek. WordNet, EuroWordNet and Global WordNet. In: *Revue Française de Linguistique Appliquee / RFLA,* Paris, France. 2002. http://home.planet.nl/~weiss075/docs/2002/rfla.pdf

Vossen, Piek. 1999. *EuroWordNet Final Report*. Deliverable D041, Work Package 0, EuroWordNet, LE2-4003, LE4-8328. Amsterdam: University of Amsterdam. http://www.vossen.info/docs/1999/D041Final.pdf

Vossen, Piek. (ed.) 2002. *EuroWordNet General Document*. http://www.hum.uva.nl/~ewn

Vossen, P. et al. 1998. *EuroWordNet Tools and Resources Report*. Deliverable D021D025, WP2, EuroWordNet, LE2-4003. http://www.vossen.info/docs/1998/D021D25.pdf

The Global WordNet Association official web site, http://globalwordnet.org

MultiWordNet official web site, http://multiwordnet.itc.it/english/home.php

Lexicographic resources

Harteveld, P. (with the assistance of L.G. De Stadler and D.C. Hauptfleisch) 1993. *Woordkeusegids: 'n Kerntesourus Van Afrikaans*. Halfweghuis: Southern Book Publishers. (*Word Choice Guide: A Core Thesaurus of Afrikaans*)

Odendal, F.F., Schoones, P.C., Swanepoel, C.J., Du Toit, S.J. and Booysen, C.M. 1994. *Handwoordeboek van die Afrikaanse Taal*. Midrand: Perskor. (*Desk Dictionary of the Afrikaans Language*)

Eksteen, Louis Cornelius. 1997. *Groot Woordeboek: Afrikaans-Engels, Engels-Afrikaans*. Kaapstad: Pharos. (*Major Dictionary*: *Afrikaans-English, English-Afrikaans*)

Taljaard, P.J. 1987. *Voorsetselwoordeboek Met Engelse Vertalings Asook Enkele Bywoorde*. Pretoria: De Jager-HAUM. (*Dictionary of Prepositions with English Translations Including Some Adverbs*)

WordNet 2.0©. Princeton University. http://wordnet.princeton.edu/.