LBK2013: A Balanced, Annotated National Corpus for Norwegian Bokmål

Rune Lain Knudsen & Ruth E. Vatvedt Fjeld

Institute of Linguistic and Nordic Studies, University of Oslo

r.l.knudsen@iln.uio.no, r.e.v.fjeld@iln.uio.no

ABSTRACT

At the Department of Linguistics and Scandinavian Studies (ILN) and the University of Oslo, the task of assembling a balanced corpus representing modern Norwegian Bokmål has reached a significant milestone. The Corpus for Bokmål Lexicography (LBK) now consists of more than 100,000,000 words. These documents have been selected based on a statistical analysis of reading habits in the general population of Norway. The documents have been subject to both manual bibliographic annotation, as well as automatic morphological annotation for each document. LBK will play a central part of a set of interconnected lexical resources, the aim of which is to provide an extensive documentation of Norwegian Bokmål that covers lexical and other linguistic/lexico-syntactic aspects. This paper presents LBK2013, a subset of LBK that we consider to be an accurate and comprehensive representation of modern written Norwegian Bokmål. A description of the corpus, as well as a number of related projects are described.

KEYWORDS: NoDaLiDa 2013, Speech and Language Technologies, Northern Europe, Corpora, Lexicography, Lexical Semantics.

Rune Lain Knudsen and Ruth Vatvedt Fjeld 2013. LBK2013: A balanced, annotated national corpus for Norwegian Bokmål. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 12–20.

1 Introduction

The LBK project was initiated in 1999 in an effort to create a corpus similar to KorpusDK, the Danish corpus that has served as the foundation for Den Danske Ordbog (DDO). In addition we wanted to use the corpus for statistical studies, which made proper balancing a key requirement. In order to account for modern Bokmål specifically, the applicable texts for inclusion in the corpus were restricted to a timespan ranging from the year 1985 and onwards. This year was chosen also to avoid digitizing earlier published texts as digitized texts became more common from around this time. This also gave a reasonable time span for modern Bokmål texts.

2 Related Work

Two balanced corpora for other languages have served as points of reference and inspiration for LBK. This section presents two corpora that have provided guidance for the decisions made for LBK: KorpusDK[1] and The British National Corpus[2].

KorpusDK is perhaps the most relevant reference point as it represents Danish, a language that has a close relationship and shares a common ancestry with Norwegian Bokmål. It has also been an integral part of projects similar to the ones currently being in development at ILN. KorpusDK consists of a total of 56 million words spanning the years 1983-2002. Each text is automatically annotated with morphological information, as well as bibliographic information and genre.

The British National Corpus (BNC) contains over 100 million words. 90% of the material is written language. The construction of BNC started in 1991 and the first official version was released in 1995. BNC has served as a main point of reference for the task of deciding upon an appropriate category distribution for written text, as it is similar to LBK2013 both in terms of size and goals. However, LBK2013 differs somewhat from BNC with regards to the chosen categories for a given text. The rationale for this is partly based on differences between the literary environments and language cultures of both countries. Due to limited finances, we also had to consider what was possible to get hold of from authors and publishers. In 1999 several writers were still afraid of their texts being misused/abused or in other ways losing some of their copyrights by allowing for inclusion into a research corpus. Over the last ten years, this fear has more or less vanished.

3 Architecture for LBK

LBK makes use of IMS Open Corpus Workbench (CWB), a widely used framework for managing and querying large text corpora (Evert and Hardie, 2011). It is available for researchers through Glossa (Nygaard et al., 2008), a web-based interface for corpora developed at the Text Laboratory, ILN. This interface enables a user to perform advanced searches within a corpus, and to specify subcorpora within which to restrict searches according to various kinds of metadata.

The documents in LBK are POS-tagged with the Oslo-Bergen tagger (Johannessen et al., 2012), and annotated with bibliographic and ethnographic annotation wherever applicable. Each text is assigned a mandatory text category: fiction, non-fiction, newspapers, subtitles or non-standardized, each of these categories divided into subcategories such as magazines, biographies, scholarly texts, etc. Each text is optionally given one or more topics such as sports,

[1] `http://ordnet.dk/korpusdk`
[2] `http://www.natcorp.ox.ac.uk`

law, medicine, ecology, music, etc. Additionally, each text is given information about the author and publishing information, wherever applicable.

LBK has received material from a large number of text suppliers over the years. One of the major bottlenecks in the workflow for LBK has thus been the handling of various document types that need to be converted to an appropriate format for the CWB framework. To enable an efficient and consistent workflow for managing the conversion and annotation of documents, a desktop application called *LBKTexts* was developed during autumn 2012. This application has simplified the workflow, increased the growth rate of LBK and secured the overall data consistency. LBKTexts is written in Java and will be made freely available under an open-source license during autumn 2013.

4 LBK2013

LBK2013 is a subset of the total amount of texts in LBK, selected with the aforementioned balance requirements in mind. The overall category distribution for LBK2013 is shown in Table 1.

Newspapers	10%
Non-fiction	45%
Fiction	35%
Subtitles	5%
Non-standardized	5%

Table 1: Overall distribution over the main categories of LBK2013

LBK was initially designed with a set of text categories, subcategories and topics, along with distributional guidelines based on observations done for similar corpora and the Norwegian Media Barometer from 2003, an annual statistical survey about the use of mass media in Norway. In 2013, this picture has changed quite drastically in some respects, especially related to internet usage. The distinction between standardized and non-standardized texts is not as easy to define anymore as much of what people read is published online without necessarily being subject to traditional proofreading. The categories and subcategories have thus been somewhat revised for LBK2013 in order to account for these changes, in particular limiting the non-standardized texts to the types of texts that without doubt can be said to be non-standardized (e.g. blogs, online forums, usenet discussions etc.).

5 Interconnectivity with a Dictionary

The need for a link between dictionaries and corpora for lexicographic work has emerged during the last 20 years, and can now be said to be the norm. A central phenomenon when analyzing corpora is the degree of word-sense ambiguity and the consequences arising from this phenomenon. Word-sense disambiguation (WSD) is thus an attractive element of a corpus analysis toolbox.

There are a number of methods one can employ when disambiguating word-senses in corpora, one popular approach being a distributional approach where the surrounding context of a word is subject to statistical analyses (cf. Kilgarriff and Tugwell (2002)). Establishing a strong link between a corpus and a dictionary should be valuable to such analyses. By using information from a dictionary as a machine-readable knowledge base, context analyses making use of e.g.

the collection of Lesk algorithms (Agirre and Edmonds, 2007) can be applied. This will also provide feedback on the consistency and completeness of the dictionary material.

By augmenting concordance results with definitions from a dictionary, users will have access to sense information for each word, including the written definition, usage examples, and so forth. In addition, users will be able to do corpus queries in which the desired subcorpora are restricted not only to morphological and bibliographic attributes, but semantic attributes as well. The dictionary will also have access to specific tokens and their context in the corpus, simplifying the task of empirical studies on dictionary lookups and related phenomena.

5.1 Establishing the Connection

A sense inventory with an index over all lemmas represented in the dictionary is made, mapping the lemma to one or more senses, each sense represented by a unique identifier either extracted from the database or generated automatically. An initial linking step is subsequently done by performing a lookup for each lemmatized word in the corpus, storing the set of possible sense id's as a special attribute for the word in question. This results in a relation from each token in the corpus to one or more senses in the dictionary whenever a match is found.

At the time of writing this paper, a test run of the linking step was done for a subset of LBK containing roughly 91.4 million words. The necessary dictionary information was extracted from Bokmålsordboka[3], a dictionary developed by ILN. A total of 59.8 million words were linked to matching dictionary entries. For these words, the average ambiguity (i.e. number of possible senses for the word) was 7.56, with a standard deviation of 1.82. There were a total of 5.93 million unambiguous words (i.e. pointing to one and only one possible sense for the word).

The ambiguous relations do provide useful lexicographic information to the user for words in the corpus matching a dictionary entry. In addition, the words that have no senses assigned indicates possible areas of study for future improvements of the dictionary. A more sophisticated approach is however needed in order to make the linked dictionary valuable for computational purposes.

5.1.1 Preliminary Disambiguation Experiments

To generalize the linking step, a sense-tagger meant to be used as an add-on for the Oslo-Bergen tagger is currently under development. Some preliminary experiments have been performed to test the functionality of the tagger. A set of sentences have been randomly extracted from LBK for the purpose of making a test set. These will be subject to manual sense-annotation using the sense inventory extracted from BOB as reference material. At the time of writing this paper, 68 sentences have been annotated manually by three annotators. A total of 907 words were assigned a total of 1254 senses. This might indicate a somewhat unneccesary degree of sense-granularity in some parts of the sense inventory. This suspicion is strenghtened by the low amount of fully agreed sense assignments done by the three annotators, as can be seen in Table 2.

The set of annotated sentences is too small to give any conclusive evidence one way or the other. However, some observations done during the annotation process was made. A large part of the cause of disagreement is due to the fine-grained sense categories. Many of the senses for a word

[3] `http://www.nob-ordbok.uio.no`

Senses	1254	100.00%
Full Agr.	552	44.01%
Major Agr.	268	21.37%
Minor Agr.	434	34.61%

Table 2: Summary of sense annotation results. **Full Agr.** is the number of senses assigned by all three annotators, **Major Agr.** is the number of senses assigned by two annotators, and **Minor Agr.** is the number of senses assigned by only one annotator.

tend to overlap, as has been observed in a number of other related projects and experiments ((Kilgarriff and Rosenzweig, 2000)). This especially holds for prepositions and auxilliary verbs. Whether or not a disambiguation at such a fine-grained level is actually neccessary has been subject to discussion in similar experiments (Palmer et al., 2006).

The sense-tagger initially assigns all possible senses for the words in the randomly selected sentences. It then walks through a series of disambiguation steps. Currently, only two disambiguation steps are actually in use: a Part-of-Speech (PoS) disambiguation, followed by an implementation of the Simplified Lesk (Kilarriff and Rosenzweig, 2000) algorithm. The PoS disambiguation removes all senses where there is a mismatch between the grammatical class that the definition belongs to, and the grammatical class assigned to the word by the Oslo-Bergen tagger. The Simplified Lesk algorithm ranks the possible remaining senses for each word by the overlap score of examples and glosses over sentence context. To increase the amount of possible overlapping words, all words subject to the overlap process are stemmed using the Snowball stemming framework[4], a set of programming libraries and tools containing improved versions of the Porter stemming algorithm for several languages. We considered lemmatizing the examples and definitions in order to do the overlap analysis on lemmas, but earlier experiments in tagging short, compressed sentences have shown that the lack of context makes automatic PoS inference unreliable. After ranking the sense candidates, all but the highest scoring candidates are discarded.

We intend to use the resulting scores from this tagger as a guideline for establishing a baseline for evaluating future versions of the sense-tagger. Currently, our test set of 68 sentences is not sufficient to make any concrete assertions. Some cautious remarks can nonetheless be made when looking at the present scores (see Table 3 for precision, recall and f-measures). Disambiguating by PoS did not improve the scores substantially, and it would be interesting to compare a larger experiment with similar experiments in other languages to see if something can be observed regarding homographs spanning more than one grammatical category. The Lesk algorithm increases the score, albeit a low one even for such a simple algorithm. The high degree of granularity for the sense inventory in use is one of the probable causes for this.

6 Interconnectivity With Other Resources

We are currently developing a framework for linking a selection of resources for lexicography and language technology. The framework will provide a communication layer that enables a given resource to query the other resources in order to supplement its own data, enabling synergetic properties to arise from the combined resources.

[4] `http://http://snowball.tartarus.org`

	None	PoS	Lesk
Precision	0.122	0.142	0.370
Recall	0.996	0.906	0.498
F-measure	0.218	0.245	0.424

Table 3: Precision, Recall and F-measures for the sense-tagger prototype under development, using no disambiguation (**None**), disambiguation by PoS (**PoS**), and further disambiguation by the Simplified Lesk algorithm (**Lesk**).

The dictionary used for the pilot project is already implicitly linked to Norsk Ordbank[5], as well as NorNet (Fjeld and Nygaard, 2009) by using common unique identifiers for senses. NorNet, a prototype for a Norwegian wordnet developed at ILN, was created by analyzing the dictionary entries in BOB, establishing a link between word-senses and synsets in NorNet and the dictionary entries in BOB. This means that the link between LBK2013 and BOB will also result in linkage between LBK2013, Norsk Ordbank and NorNet.

We wish to perform the same experiments using other lexical resources as the source of additional sense inventories. A full-scale wordnet for Norwegian Bokmål and Nynorsk is now available. This wordnet is developed by Kaldera Språkteknologi[6]. The Kaldera wordnet is based on a translation of the the Danish wordnet (Pedersen et al., 2009), and they both contain links to Princeton Wordnet (Fellbaum, 1998) via relations denoting synset-equivalence across separate wordnets. If the glossary for the Kaldera wordnet proves to be enough material for an adequate sense inventory, new possibilities for interconnectivity will arise, extending the lexical resources to cover several languages.

Glossa provides a common interface to several corpora. We plan to make use of the functionality of this interface to investigate potential candidates for interconnectivity between LBK and other corpora. The first corpus subject to such an investigation will be the Nordic Dialect Corpus. This corpus includes information such as phonetic annotation as well as transcribed audio. The aim is thus to connect tokens and sentences in LBK to spoken equivalents in the Nordic Dialect Corpus, as this information is absent in BOB.

7 Future Work

LBK2013 will be of value to a number of projects that require a balanced corpus of a substantial size. This section presents some of the remaining work to be done for LBK2013, as well as a selection of planned projects that will make use of LBK.

As shown in section 5.1.1, the relations between words in the corpus and dictionary entries need to be disambiguated further. Experiments on reducing this ambiguity automatically is in progress. We will experiment with other variants of the Lesk algorithm as well as more sophisticated algorithms. We will also attempt to make use of the metadata available in LBK2013 to see whether this can improve the disambiguation somewhat.

[5] `http://www.edd.uio.no/prosjekt/ordbanken`
[6] `http://www.kaldera.no/`

7.1 Development of a Database for Multiword Expressions

LBK will serve as the main dataset for a statistical analysis designed to assist the discovery of multi-word expressions (MWE's) and collocations. This type of analysis was done on the 40-million version of LBK in 2008 (Fjeld et al., 2010). The result of this analysis will serve as the foundation for both a lexical database designed to aid language research, and a new human-readable dictionary for MWE's and collocations. We also plan to compare the analysis done for the 40-million version with the 100-million version in order to document the benefits of enlarging a corpora. The analysis can be refined to investigate subcorpora, which will make it possible to document the phraseology of different subject fields, differences in phrases used by certain age groups, geographic locations, and more.

7.2 Lexical/Linguistic Resources

By using a large balanced corpus for lexicographic studies, hypotheses can be verified or falsified empirically. The use of corpora in lexicographic research has become the standard approach for modern lexicographic work, and it is one of the cornerstones for the continuous development of theoretical foundations for lexicography.

LBK is available as reference material for investigating existing lexical resources. It can be used as evidence for suggested revisions of existing dictionaries, such as removal of rare lemmas and/or inclusion of new lemmas based on frequency information extracted from LBK. Due to the bibliographic annotations, this can be further refined to specific category domains, for example the most frequent lemmas within the domain of law.

LBK2013 will also prove useful in the field of language standardization by providing frequency analyses of variant forms, either on a general level or for specific areas like age, sex, place of birth, even for specific sets of authors.

LBK2013 will constitute as an integral part of the BRO-project (Bokmålets og riksmålets ordbase), a collaborative effort initiated by ILN and Det Norske Akademi for Sprog og Litteratur, the aim of which is to provide an extensive lexical resource for Norwegian Bokmål based on the combined lexical resources of both parties. This will in part be based on the work done on enabling interconnectivity between a number of lexical databases at ILN. By linking data from Norwegian wordnets, the MWE database currently in development, dictionaries and LBK, we plan to create an extensive resource for Norwegian Bokmål, both machine and human readable.

LBK and the future MWE database will provide material for research and development of other lexical resources like wordnets, framenets etc. Extraction of domain knowledge should prove feasible due to the topic annotations (e.g. economy, law, medicine, computer science). We will conduct experiments on word-sense disambiguation using both statistical analyses and dictionary- and knowledge-based methods. Based on the interconnectivity framework previously described, we want to conduct experiments on relations between definitions in a dictionary and tokens in LBK, as well as information from wordnets.

Various resources useful for language technology will be made available. Frequency lists and n-gram statistics will be made available, both for the corpus as a whole and user specified subcorpora. Upon completion of the interconnectivity efforts, semantic annotations from the dictionary, wordnet and MWE will be available as source data for language technology and related research.

Acknowledgements

We would like to thank Johanne Wictorsen Kola for her meticulous work on collecting and processing texts for LBK in this important period leading up to LBK2013. We would also like to thank Kjersti Wictorsen Kola for her supervision and collaboration on the LBK material, her bibliographic knowledge, and for taking part in the sense-annotation work used in this paper. Finally we would like to extend our gratitude to the Text Laboratory at the University of Oslo for hosting LBK, providing important technical support over the years, and for developing Glossa.

References

Agirre, E. and Edmonds, P., editors (2007). *Word Sense Disambiguation - Algorithms and Applications*, chapter 5, pages 107–131. Springer.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millenium. In *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham.

Fellbaum, C., editor (1998). *WordNet - An Electronic Lexical Database*. MIT Press.

Fjeld, R. V. and Nygaard, L. (2009). NorNet - a monolingual wordnet of modern norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7 of *NEALT Proceedings Series*, pages 13–16.

Fjeld, R. V., Nygaard, L., and Bick, E. (2010). Semi-automatic retrieval of phraseological units in a corpus of modern norwegian. In *Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexicographie*, volume 25.

Johannessen, J. B., Hagen, K., Lynum, A., and Nøklestad, A. (2012). OBT+Stat: A combined rule-based and statistical tagger. In *Exploring Newpaper Language*, volume 49 of *Studies in Corpus Linguistics*, pages 51–65. John Benjamins.

Kilarriff, A. and Rosenzweig, J. (2000). English SENSEVAL: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Kilgarriff, A. and Rosenzweig, J. (2000). Framework and results for english SENSEVAL. In *Computers and the Humanities*, volume 34, pages 15–48. fd.

Kilgarriff, A. and Tugwell, D. (2002). Sketching words. In *Lexicography and Natural Language Processing*. Euralex.

Nygaard, L., Priestley, J., Nøklestad, A., and Johannessen, J. B. (2008). Glossa: a multilingual, multimodal, configurable user interface. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008.

Palmer, M., Fellbaum, C., and Dang, H. T. (2006). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. In *Natural Language Engineering*, volume 12.

Pedersen, B., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.