Extending Czech thesauri using word-formation network

Karolína Hořeňovská

Charles University, Faculty of Mathematics and Physics
horenovska@ufal.mff.cuni.cz

Abstract: In this paper, we attempt to extend existing Czech thesauri by using a word-formation network, DeriNet. Thesauri are an important resource for synonym retrieval / substitution generation but their lexical sparsity is an issue in Czech. We discuss the properties of existing thesauri and DeriNet and propose several ways of using DeriNet to extend the thesauri, such as deriving a synonym of an adverb from a synonym of corresponding adjective. We also evaluate some of our proposals.

1 Introduction

A lot of effort has been invested in creating large thesauri, of which the best known example is probably WordNet [11], followed by others such as FrameNet [2]. While these thesauri address English, there are many thesauri for other languages as well (there are e.g. WordNet versions for Arabic [4], Swedish [23], or Czech [16]). We wish to emphasize Czech WordNet since Czech is the language we currently deal with.

However, those thesauri are heavily incomplete for some languages, including the above-mentioned Czech language. This incompletness presents a problem for various NLP tasks, e.g. substitution generation as part of lexical simplification (see [20] or [19] for more detail).

On the other hand, for some languages (including Czech), a rich word-formation network is available. We propose using such network to extend existing thesauri, i.e. to discover synonymy relations between new pairs of words. Please note that while we target synonymy, as it is the only relation covered by all existing Czech thesauri, the approach would hold for any relation.

The rest of the paper is organized as follows: we briefly describe existing related work (section 2), present existing Czech thesauri (section 3) and describe the Czech word-formation network DeriNet (section 4). We then introduce several ways of combining DeriNet with thesauri to produce new relations (section 5) and evaluate the most promising of them (section 6).

2 Related work

Since thesauri are generally incomplete, there have been lots of attempts at extending them in an automated way. These attempts have included aligning multilingual resources (e.g. [22]), mining the Wikipedia ([1], [24]) or the web in general ([10]), translating English WordNet (which has been tried especially for Czech [5], even though the extension itself, to the best of our knowledge, is not publicly available), making use of word embeddings ([17], [7]) as well as by employing derivational morphology ([8], [12]).

This paper is in its nature similar to a previous attempt of extending Czech WordNet with derivational relations [14] which the authors claimed was successful. Unlike them, we use a publicly available source of derivations and we do not limit ourselves to WordNet – we try using various thesauri and compare the outcome obtained with each of them and with their combination. We also share a more thorough evaluation of the resulting pairs.

3 Existing thesauri

We are aware of five notable Czech thesauri:

- the most recent version of Czech WordNet [16], and
- a slightly divergent version of Czech WordNet [13], which lacks some synsets but contains some others which were created to enable the lexico-semantic annotation of Prague Dependency Treebank ([3]), which we refer to as WordNet (PDT);
- thesaurus formerly distributed as a part of office software LibreOffice,
- Czech Wiktionary, and
- ÚFAL thesaurus, a thesaurus developed at our department.

Both WordNet versions explicitly utilize *synsets*, each synset represents a meaning and lists literals (words or phrases) which can be used to express the meaning. Synsets might include a definition of the meaning but few have it filled.

The last three thesauri employ synsets implicitly, either by assigning a word with a set of sets of synonyms (as done in LibreOffice thesaurus and Wiktionary) or by listing sets of synonyms and including some words in more such sets (as done in our department thesaurus).

We perform our experiments both using each of the thesauri individually and using a concatenated thesaurus, i.e. an artificial thesaurus created by concatenating all synsets from each of the real thesaurus.

In our work, we do not make use of synsets. For each word, we merge its synonyms from all synsets and produce a set of its synonyms (despite the context, i.e. words which share the meaning at least in some contexts). This is partially to simplify the proof of concept, partially because senses in both WordNets are much more fine-grained than senses in other resources.

However, this step is in no way crucial. One could keep the synsets, and whenever we refer to retrieving synonym, they could first retrieve the synsets and only then retrieve the words (either from specific or all synsets). We actually expect to do this in our future work.

Some further statistics about the thesauri are provided in table 1. The concatenated line corresponds to concatenating all synsets. Please note that we only work with single-word expressions (as opposed to multi-word expressions).

4 DeriNet

DeriNet [18], [25] is a Czech word-formation network. Its nodes are Czech lexemes, i.e. lemmata, and the nodes do not have to cover all sensesl. The authors report to have decided to take a rather minimalistic approach to polysemy, and only represent a lemma with more nodes if at least one of two conditions is met: it was coincidentally derived from two different words (could be demonstrated by verb *proudit*, which is represented as a base word, though it is likely related to noun *proud* 'flow', and also as a verb derived from *udit* 'to smoke', when *proudit* refers to smoking something thoroughly), or the senses lead to different sets of derived words (i.e. verb *stát* 'to stand, to melt away').

The directed edges then represent the fact that one word is derived from the other one. The edges should be taken as implicative, some derivations might not be captured in DeriNet (yet). They are discovered using a variety of methods, including manual deduction, rule-based automated processes and machine learning; many of them were also taken from the MorfFlex CZ morphological dictionary [6]. All discovered edges are manually confirmed before being added to the network.

By the authors' design decision, no word is allowed to have more than one parent, which simplifies the structure and could be justified by low occurence of compounds in Czech. Even though only one parent is allowed, recent versions of DeriNet allow for an indication of being a compound in part of speech specification.

The then current version of the network (1.7) contains $1,027,655$ nodes, though only some of the nodes are supported by corpus evidence (when compared to SYN v4 version of Czech National Corpus [9], we found out that as many as $591,486$ nodes (i.e. more than a half) do not occur in the corpus). For the first version, only words which occured at least twice in a SYN subcorpus of Czech National Corpus (and fullfiled a few other conditions) were inserted in the network; this condition does not hold for lemmata inserted from MorfFlex CZ dictionary.

Of all nodes, $104,563$ (approx. $10\,\%$) are isolated, i.e. they are not connected with any other node.

Except for the parent and part of speech, there is no further annotation, i.e. one cannot learn for example that the derived noun is agent noun of the base verb. DeriNet format is therefore farily simple: it gives node ID, its lemma and technical lemma (which contains some additional details such as sense disambiguation), its part of speech (perhaps with the above-mentioned indication of being a compound) and its parent's ID (if the node has a parent).

5 Proposed thesauri extensions

We propose the following principle of discovering new word relations:

1. Find a non-root node A (i.e. a node which has a parent).
2. Get A's parent, B.
3. Retrieve B's synonyms using the existing thesauri.
4. Find all nodes C which correspond to the retrieved synonyms.
5. For each C, check if it has a child D which shares requested features with A.
6. Declare A and D a related word pair.

This outline does not specify how to deal with the situation when more than one D exist (share given features with A) for single C. In our experiments, we opted for choosing neither (i.e. skipping the whole C subtree) but one could also develop strategies to select the best D or generate more pairs for A from single C.

It should be noted that due to this decision, discovering synonymous word pairs is not symmetric, that is, a word pair might be discovered when starting from one word, but not when starting with the other one.

We actually suggest further constraining all of A, B, C and D to improve the reliability of the discovered relations, i.e. by constraining their part of speech. While part of speech is the only feature available in DeriNet itself, we can use e.g. MorfFlex CZ dictionary or MorphoDiTa tool for morphological analysis [21] to enable more features.

One could be tempted to only search for those non-root nodes A which are not covered by any thesauri, the reasoning being that such nodes already have their synonyms in the thesauri. However, thesauri entries for individual words are often incomplete and the outlined process could still find new synonyms for node A, even if node A is present in a thesaurus. Furthemore, considering only nodes A which are not covered by thesauri could lead to a decrease in number of retrieved pairs after adding a new thesaurus as some nodes could be newly skipped. We therefore do not constrain node A on its presence/absence in the thesauri.

Thesaurus	Synsets	MWE	SWE	Pairs	1+ synonyms	N	A	V	D
Department	18211	9	21382	40215	21382	46%	27%	23%	4%
LibreOffice	39460	17310	32085	118243	32085	42%	26%	24%	4%
WordNet (PDT)	23094	7696	17599	18015	10962	63%	9%	27%	1%
WordNet	28459	10846	21275	20067	14095	65%	9%	24%	1%
Wiktionary	43138	1038	40319	16657	11268	62%	20%	15%	3%
Concatenated	152362	28463	69064	162709	43516	54%	22%	20%	3%

Table 1: Basic statistics of thesauri: numbers of synsets, multi-/single-word expressions, unique synonym pairs (if a synonym pair can be retrieved from more than one synset, it is only counted once), words with at least 1 synonym; POS distribution for nouns, adjectives, verbs, adverbs (the distribution need not sum up to 1 because some thesauri contain also other POS such as prepositions)

While we describe the process as deriving synonyms by using the synonymy relation of parent nodes, the parent-child relation is not crucial. One could reword the process e.g. with finding A's child B and C's parent D, or with finding A's grand-parent B (and C's grand-child D). However, by the nature of thesauri creation, we expect them to contain the base word rather than the derived word (though the direction of the derivation is sometimes ambigous). Longer distance relations, on the other hand, are more likely to introduce noise.

Having introduced the basic principle, we suggest specific approaches to relation derivation. We assume the access to richer morphological annotation, e.g. by using the MorphoDiTa tool.

5.1 Deriving adverb synonyms using adjectives

Adverbs are often derived from adjectives and while adverb ratio in thesauri is close to corpus ratio (approx. 2.6 %-3.2 % of content words as measured using Czech National Corpus syn v4), we often ran into issues with them in our text simplification experiments.

We suggest constraining nodes A and D to be adverbs and nodes B and C to be adjectives.

5.2 Deriving feminine forms from masculine

In Czech, some words come in different forms for men (or males generally) and women. This in particular holds for roles in relationships and for agent nouns, e.g. there is *učitel* 'teacher (man)' and *učitelka* 'teacher (woman)'.

Thesauri usually only cover the masculine variants, both because they are usually the default and because native speakers can infer the feminine variant (still, some language knowledge is required, e.g. there is *učitelka* to *učitel* but *ministryně* to *ministr* 'minister', not **ministrka*).

We suggest constraining nodes A and D to be nouns having feminine gender and nodes B and C to be nouns having masculine gender.

5.3 Deriving possesive adjectives from nouns

Similarly to omitting feminine forms, thesauri generally do not cover possesive adjectives since they can be easilly inferred from the corresponding noun.

We suggest constraining nodes A and D to be possesive adjectives and nodes B and C to be nouns.

5.4 Deriving verbs using verbs of opposite aspect

Thesauri differ in treating verb aspects, and often thesauri are not consistent even internally. Sometimes the verb of opposite aspect is listed as synonym, sometimes both aspects form their own synset, sometimes the other aspect is completely missing.

We suggest constraining nodes A and D to have matching aspect, nodes B and C to have matching aspect, and nodes A and B to have opposite aspect.

The issue with this suggestion is that aspect information is neither present within MorfFlex CZ dictionary nor provided by MorpohDiTa. We believe, however, that annotation from Czech National Corpus (e.g. the beforementioned version syn 4) [9], which is enriched with aspect annotation, could be used.

6 Evaluation

We tried generating synonym pairs from adverbs using adjectives, feminine forms using masculine forms and possesive adjectives using nouns (see section 5 for more detail).

The generation procedure was carried out for each of the thesauri individually and also for the result of concatenating all the thesauri together.

Our results are reported in table 2. We report the number of obtained pairs for each of the thesauri as well as for the concatenated thesaurus. The reported numbers are after symmetrization, i.e. after expanding any pair A-D into both A-D and D-A. The actual numbers of discovered pairs are usually 1.6-1.9 times greater as most pairs (but not all of them) are discovered in both directions. These ratios seem to slightly correlate with the selected strategy (symmetrization is of greatest help when finding feminine variants).

When evaluating a specific thesaurus, we can discover a synonym pair which is actually present in some other thesaurus. Whenever this happens, we consider such pair

	Department	LibreOffice	WordNet (PDT)	WordNet	Wiktionary	Concatenated
			Adverbs via adjectives			
Obtained	21,498	26,293	731	952	318	34,766
Confirmed	1,241	1,481	58	97	11	1639
Precision	0.77 when accepted by 2+ annotators			0.47 when accepted by all annotators		
			Feminine via masculine			
Obtained	1,392	2,312	754	760	973	3,574
Confirmed	72	89	32	36	99	129
Precision	0.82 when accepted by 2+ annotators			0.48 when accepted by all annotators		
			Possesive adjectives via nouns			
Obtained	2,292	4,329	1,089	1,094	2,818	7,512
Confirmed	0	0	0	0	0	0
Precision	0.84 when accepted by 2+ annotators			0.61 when accepted by all annotators		

Table 2: Results of our synonym pair generation. We report the number of pairs obtained using the thesauri, number of pairs confirmed by existing thesauri and human-rated precision on a sample of 100 pairs.

a *confirmed* one. We do not evaluate it further and expect that the pair is correctly derived and synonymous.

For each strategy, we sampled 100 non-confirmed pairs and asked 4 annotators to annotate them as either synonym, antonym or unrelated. The annotators were of varying gender and age, though all of them have obtained a university degree during their life. Asking annotators to distinguish antonyms from unrelated pairs was done based on our informal result analysis, which revealed antonym pairs do occur.

In 5 cases, the annotators admitted they did not know, in a few other, they noted they were not really sure. In all cases, a very infrequent word was involved. We treat *I don’t know* as *unrelated* when reporting precision and inter-annotator agreement. We treat answers marked with *not sure* in the same way as unmarked.

The inter-annotator agreement (Fleiss’ kappa) was 0.47 and it slightly varied over the strategies (0.47, 0.42 and 0.50, respectively). These numbers might seem low but it is important to keep in mind that most answers were *synonym*, hence this answer had a great probability, and therefore any disagreement on other answers had a big impact.

Examples of correctly discovered pairs (pairs annotated as synonyms by all four annotators) are given in table 3. There were 9 pairs marked as antonyms by all annotators (out of 28 marked as antonyms by at least one annotator), they are all listed in table 4. In all 9 cases, the pair was derived using a synonymy relation from LibreOffice thesaurus, when either two antonyms were suggested as synonyms to the same word, e.g. both *tlouštík* ’fatty’ and *hubeňour* ’thin man’ to *tlust’och’* ’fatty’, or when an antonym was suggested directly as e.g. *inkompatibilní* ’incompatible’ to *kompatibilní* ’compatible’. While these pairs are not synonymous, their existence should not be used to decline the principle. On the contrary, should the thesaurus pairs be correctly marked as antonyms, we would correctly derive antonymous pairs using our method.

vodpovědně	*spolehlivě*	dependably
hanebně	*bezcharakterně*	unscrupulously
vyzvědačka	*špehounka*	she-spy
čarodějnice	*divotvorkyně*	witch
surovcův	*krut’asův*	bully’s
maršálkův	*maršálův*	marshal’s

Table 3: Examples of correctly discovered pairs

Finally, there were 14 pairs annotated as unrelated by all annotators. They are listed in table 5. In some cases (1, 4, 5), there is some evidence that the words can share a meaning but at least one of the words is associated with another meaning so strongly that the annotators probably did not realize the meaning could be the same.

Some pairs (2, 3, 10, 11, 13) come from a synset in thesauri, even though we could not find any other evidence that these pairs really could share the meaning.

Other pairs (6, 14) occur because of insufficiencies in the derivational process. While the base words are synonyms, the derived words are of distinct genders. These pairs could be prevented by constraining the suggested pairs more carefully.

Case 9 is quite similar. Both words *strůjce* ’creator’ and *otec* ’father’ could refer to a creator (author) and *strůjkyně* ’she-creator’ is a feminine variant of *strůjce*. However, while the word *otčina* ’fatherland’ is directly derived from *otec* and is feminine, it does not in any way refer to she-father. This could be prevented with more detailed annotations in the word-formation network.

There are cases (7, 8) when, despite the principle proving good, derived words are not really perceived synonymous, even though the base words could be. For example, both words *kůň* ’horse’ and *osel* ’donkey’ could be used to refer to a dumb person but their feminine variants are not used in that way (even though in theory they could be).

The last case, 12, is special in many ways. The word *zároveň* ’at the same time, simultaneously’ is reported to be derived from word *rovný* ’straight’, which might seem surprising. The pair is further derived from thesauri pair

pokřiveně	distorted-ly	*rovno*	straight
povšechně	in general	*konkrétně*	specifically
různě	diversely, differently	*identicky*	identically
mladice	young woman	*stařice*	old woman
tlust’oška	fat woman	*hubeňourka*	thin woman
živelně	elementally, unrestrainedly	*organizovaně*	organized-ly
jemně	softly, lightly	*pikantně*	spicy, zesty
inkompatibilně	incompatibly	*kompatibilně*	compatibly
bezcitně	heartlessly	*vřele*	heartily

Table 4: Pairs marked as antonymous by all annotators

1	*hospodáříčkův*	*raráškův*
2	*jedlice*	*nájemnice*
3	*jezdkyně*	*běžkyně*
4	*koňův*	*imbecilův*
5	*léčitelův*	*věštcův*
6	*nešičin*	*amatérův*
7	*oslice*	*konice*
8	*radikálně*	*zleva*
9	*strůjkyně*	*otčina*
10	*surovcův*	*katův*
11	*vinařka*	*hornice*
12	*zároveň*	*zakrouceně*
13	*čile*	*umně*
14	*št’ouřin*	*kutilův*

Table 5: Pairs marked as unrelated by all annotators

rovný ’straight’ – *zakroucený* ’tortuous’, which is rather antonymous.

We do not provide detailed report on pairs annotated differently by different annotators, though we have examined them too. In most cases, some evidence of shared meaning exist but some of the annotators did not consider the words synonymous.

Following from the above analysis, more than half of unrelated pairs is not less related that their base word counterparts. These pairs do not contradict our method, they only evidence the necessity of both checking thesauri quality and being careful about the synonymy itself as it is perceived differently by different people.

There are cases when our method fails to filter out non-synonymous derived pairs. This could be improved both by better filtering during the inference process and by having better annotation in the word-formation network.

7 Conclusion

We have presented a method of deriving new synonym pairs using existing thesauri and word-formation network, we have suggested several strategies to do the actual derivation and we have evaluated some of them.

Our evaluation revealed that about half of derived synonym pairs are really perceived synonymous by all of our human annotators and around 80 % are perceived synonymous by at least two of them. The erroneous word pairs are caused by two distinct factors. First, there are errors in the thesauri synsets: unrelated, or even antonymous, words are occasionally marked as synonymous. Second, there are limitations of our method, where the derived words are not synonymous, despite being derived from synonymous base words.

Some of the limitations could be overcome by better filtering within our method or by more detailed annotations in the word-formation network. The latter has become available soon after we carried out our experiments, as DeriNet 2.0 has been released. This version has a more detailed annotation of both nodes (e.g. noun gender) and edges (the purpose of derivation is annotated, e.g. diminutivization), and we expect this version to be helpful in future experiments.

We also plan to try using Derivancze [15] (which also includes derivation annotations) instead of DeriNet as the word-formation network and see if it helps to improve our results.

Overall, we consider our results good because they suggest that thesauri authors can focus on capturing the relations between the base words and NLP applications can still make good use of those thesauri even for derived words.

Acknowledgement

This work has been supported by the grant No. 1704218 of the Grant Agency of Charles University. It has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). The research was also partially supported by SVV project number 260 453.

References

[1] Musa Alkhalifa and Horacio Rodríguez. Automatically extending NE coverage of Arabic WordNet using Wikipedia. In *Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco*, 2009.

[2] Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[3] Eduard Bejček, Petra Hoffmannová, Martin Holub, Marie Hučínová, Pavel Pecina, Pavel Straňák, Pavel Šidák, and Jan Hajič. Lexico-semantic annotation of PDT using Czech WordNet, 2011. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[4] William Black, Sabri Elkateb, and Piek Vossen. Introducing the Arabic WordNet project. In *In Proceedings of the third International WordNet Conference (GWC-06*. Citeseer, 2006.

[5] Marek Blahuš. Extending Czech WordNet using a bilingual dictionary. Master's thesis, Faculty of Informatics, Masaryk University, 2011.

[6] Jan Hajič and Jaroslava Hlaváčová. MorfFlex CZ, 2013. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[7] Jugal Kalita et al. Enhancing automatic WordNet construction using word embeddings. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 30–34, 2016.

[8] Svetla Koeva, Cvetana Krstev, and Duško Vitas. Morpho-semantic relations in WordNet–a case study for two Slavic languages. In *Global wordnet conference*, pages 239–253. University of Szeged, Department of Informatics, 2008.

[9] Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Zasina. SYN v4: large corpus of written Czech, 2016. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[10] Robert Meusel, Mathias Niepert, Kai Eckert, and Heiner Stuckenschmidt. Thesaurus extension using web search engines. In *International Conference on Asian Digital Libraries*, pages 198–207. Springer, 2010.

[11] George A Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[12] Verginica Barbu Mititelu. Adding morpho-semantic relations to the Romanian WordNet. In *LREC*, pages 2596–2601, 2012.

[13] Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. Czech WordNet 1.9 PDT, 2011. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[14] Karel Pala and Dana Hlaváčková. Derivational relations in Czech WordNet. In *Proceedings of the workshop on balto-slavonic natural language processing: Information extraction and enabling technologies*, pages 75–81. Association for Computational Linguistics, 2007.

[15] Karel Pala and Pavel Šmerk. Derivancze—derivational analyzer of Czech. In *International conference on text, speech, and dialogue*, pages 515–523. Springer, 2015.

[16] Karel Pala and Pavel Smrž. Building Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88, 2004.

[17] Heidi Sand, Erik Velldal, and Lilja Øvrelid. WordNet extension via word embeddings: Experiments on the Norwegian WordNet. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 298–302, 2017.

[18] Magda Ševčíková and Zdeněk Žabokrtský. Word-formation network for Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

[19] Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70, 2014.

[20] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355. Association for Computational Linguistics, 2012.

[21] Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[22] Lonneke Van der Plas and Jörg Tiedemann. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 866–873. Association for Computational Linguistics, 2006.

[23] Ake Viberg, Kerstin Lindmark, Ann Lindvall, and Ingmarie Mellenius. The Swedish WordNet project. In *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002: Copenhagen, Denmark, August 13-17, 2002*, pages 407–412, 2003.

[24] Ichiro Yamada, Jong-Hoon Oh, Chikara Hashimoto, Kentaro Torisawa, Jun'ichi Kazama, Stijn De Saeger, and Takuya Kawada. Extending wordnet with hypernyms and siblings acquired from Wikipedia. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 874–882, 2011.

[25] Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1307–1314, 2016.