Extracting a Semantic Database with Syntactic Relations for Finnish to Boost Resources for Endangered Uralic Languages

Mika Hämäläinen

Department of Digital Humanities, University of Helsinki
mika.hamalainen@helsinki.fi,
WWW home page: https://mikakalevi.com/

Abstract. This paper introduces the second version of SemFi, a semantic database for Finnish with syntactic relations. The previous version of SemFi has been used in poem generation, and thus it has application area in NLG applications. In addition to extending SemFi, this paper describes and evaluates its translation into four endangered Uralic languages, Skolt Sami, Erzya, Moksha and Komi-Zyrian, all of which are greatly under-resourced. The translated dataset is known as SemUr.

Keywords: semantics, endangered languages, Finnish, Skolt Sami, Erzya, Moksha, Komi-Zyrian

1 Introduction

Endangered Uralic languages suffer from a lack of computational resources needed for statistical and neural approaches to natural language processing. A great deal of NLP work in the recent years for these languages has been focusing on rule-based systems, such as FST (finite-state transducer) morphology and RBMT (rule-based machine translation) and lexicographic work in the Giellatekno infrastructure [13].

The lack of digital resources does not come as a surprise when the languages in interest vary from severely endangered Skolt Sami with around 300 native speakers to definitely endangered Komi-Zyrian with a little over 200.000 native speakers [12].

This paper focuses on a subset of Uralic languages: Skolt Sami, Erzya, Moksha and Komi-Zyrian. The reason for choosing these languages is that a recent research in combining multilingual lexicographical resources for these same four languages [7] identified a need of making a semantic distinctions in the case of polysemy in order to achieve better results in combining these dictionaries. In other words, a Skolt Sami word *bliin* can be translated into Finnish as *levy* (disc) or *lettu* (pancake). When combining this entry with the Erzya word *пластинка* which can only mean disc, one has to be able to differentiate polysemous dictionary entries from synonymous ones through the majority language.

In this paper, we build a semantic database (SemFi) for Finnish automatically. The database consists of words which are linked to each other based on the strength of syntactic relations observed in a large, syntactically parsed corpus. Such a database can capture a multitude of semantic information, such as the actions a nouns can perform (subject relation), the attributes a noun has (adjective attributes) and the manner in which actions can be performed (adverb to verb relation).

Furthermore, the Finnish database is translated into the four endangered languages under study. These four databases are known as SemUr. The databases built in this paper have been released online to promote the resources available for these languages.

2 Related Work

The semantic knowledge of the endangered Uralic languages is limited to semantic tags in the Giellatekno dictionaries. These tags are by no means complete neither do they try to model the semantics in an accurate fashion, since their sole purpose is to serve in CALL (computer-assisted language learning) applications [2] In other words, there is a need for projecting semantic knowledge from a majority language with high resources.

For Finnish, the freely available semantic resources consist of FinWordNet and FinFrameNet [11] which are direct translations of their English counterparts. The problem of these resources is that they capture only a small part of the language and they are culturally towards the English speaking world as they are translated from the English resources. This is problematic especially in the case of Uralic languages which are culturally closer to the Finnish and Russian speaking worlds due to their geographical location. Therefore building on natively Finnish or Russian resources is a better mirror to the conceptual space of the endangered Uralic languages.

There are pre-trained word2vec models and other similar distributional semantics models available for Finnish [3], however previous research [4] has shown that a syntactically aware semantic database can be used in tasks ranging from semantics to pragmatics (such as metaphor generation) in a novel way due to the additional syntactic information not present in a word2vec model.

A large-scale FinnONTO project [9] consists of ontologies for Finnish built with the semantic web ideology. Multiple different ontologies have been developed for Finnish as a part of the FinnONTO project ranging from specific topics such as literature or health to core ontologies which are not specific to one field or theme.

The database built in this paper is an extension to an existing syntactically aware semantic database for Finnish called SemFi [4]. As pointed out in their paper, SemFi suffers from the limited number of syntactic relations that have been incorporated into the database. This limitation makes its use difficult for our needs, and thus we have to build upon it.

While SemFi has been previously used in the challenging AI task of poem generation [4], a similarly built database for English [1] has also been used in computational creativity. The English database was used as a part of slogan generation. This shows that the dataset presented in this paper has applicability in solving hard AI problems such as creativity.

3 Building the Finnish Semantic Database

The semantic database can be seen as a network that consists of lemmatized words with their part-of-speech tags. These words are connected to each other by the syntactic relations observed in a corpus. Each relation stores also the strength of the relation. Two strengths are recorded in the database: the absolute frequency of the co-occurrence of the two words given the relation and the relative frequency of the co-occurrence over all of the words linking to the head word with the same syntactic relation.

3.1 Extracting the Data

We build the database based on the syntactic bigram data of the Finnish Internet ParseBank [10]. These bigrams differ from the regular bigrams in such a way that the words are not necessarily each other's immediate neighbors in the text, but they are connected to one another by a syntactic arch. The data consists of internet text crawled as a part of the Common Crawl initiative. These texts have been automatically parsed with the Finnish Dependency Parser [8].

For both of the words in the bigram the word form, lemma, part-of-speech and morphological reading is given. The following example shows two bigrams from the Finnish Internet ParseBank data:

1. ovat ovat/ovat/V/PRS_Pl3|VOICE_Act|TENSE_Prs|MOOD_Ind| OTHER_UNK/ROOT/0 ,/,/Punct/_/punct/1 4
2. soitella soitella/soitella/V/NUM_Sg|CASE_Lat|VOICE_Act|INF_Inf1 /xcomp/0 koiraa/koira/N/NUM_Sg|CASE_Par/dobj/1 3

As noted in earlier research [6], this data consists of a multitude of parsing errors, non-words consisting of erroneous characters due to wrong encoding, incorrectly tagged or lemmatized words and so on. For instance, in the example 1 above the verb form *ovat* is incorrectly lemmaitzed to *ovat*, whereas the correct lemma would be *olla*.

As an initial filtering step, we list all the part-of-speech tags and names of syntactic relations that occur more than 1000 times in the corpus. This is because even these can have noise, mostly due the fact that the dataset separates information with slashes (/). If the word itself has a slash, e.g. a url, this will render the data effectively unparseable. We go through the list of the frequent part-of-speech tags and relation names manually to further filter out noise. Only these parts-of-speech and relations will be recorded in SemFi.

Finnish has a tendency of forming new words with compounding this means that when in English words such as *gas station* or *Ministry of Foreign Affairs* are either formed by two words written separately or with a prepositional structure, in Finnish these words are written together *huoltoasema* and *ulkoasiainministeriö*. Compound words are marked with a pipe symbol (|) in the ParseBank Data, but oftentimes there is noise in the compounds recorded in the dataset. If the part-of-speech of the compound is of a closed class or an adverb, we filter it out from the data. This is done because words of these parts-of-speech don't typically form compounds and thus compounds of this kind are mostly noise. Nevertheless, for the compounds that were acceptable, we record a value of 1 in the compound column of SemFi.

To further remove the noise, we check all the words with Omorfi [14], which is an FST based Finnish morphological analyzer. Firstly, for every word in the ParseBank data, we check whether it is lemmatized correctly and, secondly, that the part-of-speech matches the one output by Omorfi. If either of these fail, the word is not recorded in the database. This will effectively remove non-words, encoding errors and morphological parsing errors. For compounds, we only check the last word of the compound which is the one that determines the part-of-speech of the whole compound and is the only morphosyntactically inflecting part of the whole compound.

Because Omorfi is a fully rule-based system, we can trust its accuracy. However, this accuracy does not come without a trade-off. A great many neologisms such as *photoshopata* (to photoshop) are not recognized by Omorfi and thus get removed from our semantic database. However, the dictionaries of Uralic languages do not cover the most modern words at any rate, so for our purposes this trade-off is acceptable to achieve a higher accuracy in the produced database.

3.2 The Resulting Database

The structure of the SemFi database is presented in Figure 1. The database consists of two tables: *words* and *relations*. These tables are connected by the two foreign keys in *relations* referencing to *words*.

The *words* table records each word that has appeared in the corpus after the filtering steps and that has been connected to at least one word. The frequencies are calculated based on the frequencies of all of the relations the word has in SemFi. The relative frequency is the frequency divided by the sum of frequencies of all of the words in SemFi. The compound value is 1 for words that were marked as compounds in the original corpus, and 0 otherwise. It's important to note that in the case of SemUr, this cell indicates whether the word is a multi-word expression.

The *relations* table connects two words together by a syntactic relation indicated in by *relation_name*. The frequencies show the number of times these two words have co-occurred with this particular relation. Again, the relative frequency is the frequency divided by the sum of all the frequencies of where the *word1* and *relation_name* is the same. In other words, it indicates the probability of *word2* given *word1* and *relation_name*. In addition to the relative

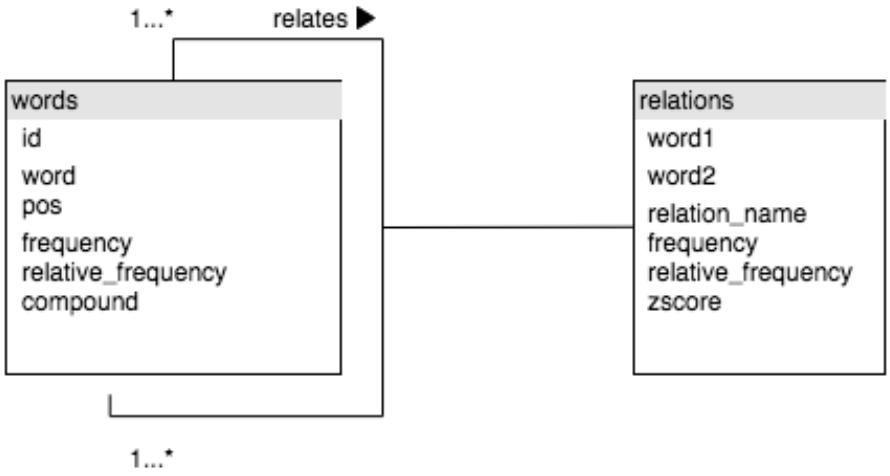

Fig. 1. A diagram of SemFi

frequency, z-score is calculated in a similar fashion[1]. In case the z-score returned a NaN value, this value is recorded as 0 in SemFi.

	N	V	A	Adv	Pron	C	Interj	Num	Adp	*Total*
Count	1 400 107	27 055	124 610	3 916	58	37	446	82	250	*1 556 561*

Table 1. Number of words in SemFi in each part-of-speech category

Table 1 shows the total number of unique words in SemFi and their distribution in different parts-of-speech. The overwhelming number of nouns in relation to other parts-of-speech is partially explained by the way Finnish forms new words by compounding. These words are interconnected by the total amount of 62 450 043 relations recorded in SemFi.

4 Projecting the Relations to the Endangered Languages

This section explains the creation of SemUr which a collection of four databases translated from SemFi for each endangered language in question. SemUr is produced by dictionary translation.

We use the multilingual Giellatekno dictionaries distributed as XML dumps through the Online Dictionary for Uralic Languages [5] for Skolt Sami, Erzya,

[1] Z-scores are calculated by using SciPy

Moksha and Komi-Zyrian as our starting point. These dictionaries are multilingual in the sense that each one has the dictionary entries in the respective minority language. Underneath each entry, there are translations to other languages. Usually, at least a translation in Finnish is provided, but it is common to have translations to other languages as well such as English and Russian in particular.

In theory, the structure of these dictionaries marks polysemy by dividing translations into multiple meaning groups. Polysemy annotation of this nature would be useful when using these dictionaries to translate SemFi, but in practice previous research using these dictionaries [7] has shown that the polysemy annotation has, for most part, been ignored by the editors of the dictionaries and thus its use would not make too big an improvement.

	Skolt-Sami	Komi-Zyrian	Erzya	Moksha
Finnish words	29 568	15 777	12 215	15 321

Table 2. Number of Finnish translations in each dictionary

Table 2 shows the number of unique Finnish translations for each language. It is evident by the size of the dictionaries that the SemUr databases will be considerably smaller than SemFi. Yet, it is worth noting that the dictionaries are rather extensive given that the languages in question are endangered and only Skolt Sami is spoken in Finland while the rest are spoken in different parts of Russia.

Even though the dictionaries follow an XML structure, they are not free of noise. Each dictionary has been edited by multiple different people during different time periods, which clearly shows as an inconsistency in the style in which the dictionary entries have been introduced into the dictionaries. The Finnish translations can have notes in brackets, multiple translations separated by comma, enumeration of translations, and question marks indicating that further check is needed. For our purposes, we remove all these additional annotations so that only one single unannotated translation is left.

The actual translation of SemFi is done so that each word recorded in SemFi is checked in a minority language dictionary for existence by its lemma and part-of-speech. If no translation is found, the word is removed, in case there is a translation available, the first matching word is used to translate the Finnish word. Word frequencies are counted again by what is left in the translated database, so that the relative frequency is still relative to SemFi. The only structural difference is that the *compound* field is now used to indicate a multi-word expression. This is because the dictionaries do not indicate whether a word is a compound word, but they have translations into multi-word expressions, which are absent in SemFi.

Table 3 shows the number of unique words in SemUr databases for each language. The Komi-Zyrian database has no conjunctions and the Moksha one

	N	V	A	Adv	Pron	C	Interj	Num	Adp	*Total*
Skolt Sami	5 004	2 356	1 012	503	16	11	7	15	62	*8 986*
Komi-Zyrian	3 236	1 116	673	173	16	0	4	33	22	*5 273*
Erzya	3 400	1 497	236	78	13	4	8	24	12	*5 272*
Moksha	1 678	2 394	716	2	0	4	0	0	0	*4 794*

Table 3. Number of words in SemUr in each part-of-speech category

no pronouns, interjections, adpositions or numerals. This is because the XML dictionary dumps for these languages did not contain any words in those parts-of-speech. The total number of words in SemUr is lower than the number of available translations, the reason for this is discussed in the *Results and Evaluation* section.

5 Results and Evaluation

In this part, we will conduct evaluation on the SemUr databases. We will shed more light into why only a fraction of the translations provided in the dictionaries ended up in SemUr. What type of words were not translated from SemFi and what type of words were not present in SemFi while present in the dictionaries of the endangered languages. In addition to this evaluation, we conduct evaluation of the quality of the translations by the help of human annotators.

5.1 Overlap of the Dictionaries

As noted in the previous section, only a small part of the words in SemFi were translated into the endangered languages. In addition to that, the original dictionaries were bigger in word coverage than the translated SemUr databases. In this section, we present some initial analysis on the overlapping words and the ones that were not translated.

Figure 2 indicates that there is a huge amount of unique vocabulary in all of the XML dictionaries that is only covered in one dictionary. The biggest single overlap (3999 words) is between the Skolt Sami dictionary and SemFi, but following that the second largest overlap (2240 words) is at the intersection of all of the dictionaries and SemFi. However, the largest numbers of words are in the petals of the diagram. Next we will take a brief look into the words that are covered in all of the datasets and the ones that are unique to one dataset.

The intersection consists of only of nouns, verbs and adjectives. The shared vocabulary consists mainly of fundamental concepts such as colors (*green, to grey*), emotions (*to be disappointed, to mourn*), words referring to mental processes (*to forget, to hope*), nature (*squirrel, stallion*), human relations (*father in law, slave*) and so on. An interesting remark, which highlights the importance of the hypothesis presented earlier about building on a culturally similar majority language, is that culturally important concepts such as *skiing*, religious concepts

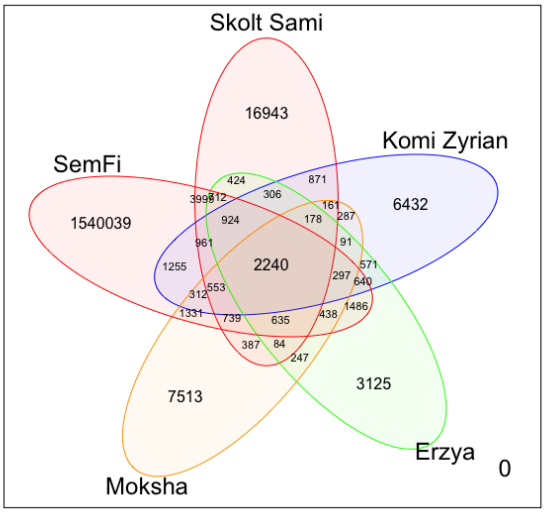

Fig. 2. A Venn diagram showing the overlap of SemFi and the Giellatekno dictionaries

such as *church* and *sin*, and concepts related to Russia such as *ruble* and *boyar* are present in all of the datasets.

The Skolt Sami words that were not used in translation include a great many multi-word expressions such as *varttunut vasa elokuulla* (a calf that has grown up in August) and compound nouns such as *oinaantalja* (coat of ram). Also, many frequent morphologically derived words have not been used in the translation such as the noun *pihkaantuminen* (the act of becoming stained with pitch) form the verb *pihkaantua* (to become stained with pitch) and the adjective *kääpäinen* (having polypores) from the noun *kääpä* (polypore). Words that are used prefixally as a part of a compound word have not been used in the translation either such as *myöhäis-* (late).

The Komi-Zyrian dictionary words that have not ended up in SemUr, have mainly the same reasons as in the case of Skolt Sami. Words translated with multiple words such as *loimitukin kiristäjä* (a tightener of the fore beam of a loom) and compounds like *syyssulkasato* (autumn molting of feathers) are frequent in the list of non-translated words. There are also some mismatches in parts-of-speech in relation to those in SemFi, for example quantifier (Qnt) and particle (Pcle) are used. An example of a quantifier would be *kolmisin* (the three of us/you/them) and a particle *yhdessä* (together). These words have not even been considered in the translation step, because of the requirement of the same part-of-speech in SemFi and the translation.

What comes to the Erzya dictionary, while the same reasons as in the case of Skolt Sami and Komi-Zyrian seem to be prevalent, the dictionary also has a great many translations that are, in fact, example sentences in Erzya followed

by their Finnish translation. An example of this phenomenon is *начко пенгтне мознить kosteat puut kytevät* (wet trees smolder). This is an example of the fact that these dictionaries do not always follow the structure of the Giellatekno XML, which has a separate element (<xt>) for example sentences.

The unused Moksha translations mainly follow what has been discovered with the other languages. Interestingly the dictionary contains a myriad of frequentative verb forms, such as *päällystellä* (to coat casually) from *päällystää* (to coat) and *siivoilla* (to clean casually) from *siivota* (to clean).

5.2 Evaluation of the Translations

In order to conduct evaluation on the translations in the SemUr databases, we sample 20 words at random out of the 300 most frequent words in the database of each language. For all of these words, we take the top 2 most frequent syntactic relations and for each relation 5 words connected by that relation. All in all, we have 200 word1, relation, word2 triplets for all 4 languages to evaluate. These triplets are evaluated by linguists knowledgeable in these languages in terms of two evaluation questions.

1. Is the relation possible for the word1?
2. Can the two words be related to each other with the given relation?

The purpose of these questions is firstly to evaluate the amount of noise in the relations and secondly evaluate how accurately the word-level translation worked. In case of a negative answer, the evaluators were asked to provide additional comment on why they considered the triplet wrong.

In the end, every language was evaluated by one person, except for Moksha which was not evaluated due to not finding any suitable evaluator with enough time to dedicate on the matter. The Skolt Sami evaluator went through only 150 out of the 200 triplets. Recruiting evaluators with enough linguistic background knowledge and a good command on the language is difficult in the case of endangered languages.

	Q1 - yes	Q1 - no	Q2 - yes	Q2 - no
Skolt Sami	93.3%	6.7%	71.5%	28.5%
Komi-Zyrian	92%	8%	60%	40%
Erzya	92.5%	7.5%	65.5%	34.5%

Table 4. Results from the human evaluators

Table 4 shows the quantitative results based on the evaluators' judgments. Most of the time, the syntactic relation has been considered possible by the evaluators. The reasons for the wrong relation had mainly to do with the valency of the word1. For instance, verbs that are transitive in Finnish had been translated with an intransitive verb in the other languages. An example of this is the Finnish *ajaa* (to drive) translated in Komi-Zyrian as *исковтны*.

In Erzya, the Finnish word *toivoa* had been translated by *мель максомс* which literally translates into *to give desire*. This verb can have direct objects in Finnish, whereas in Erzya the multi-word expression already has a direct object and cannot thus take another one.

As for the second question of the two words connecting to each other by the relation, most of the errors are due to semantic incompatibility. Although, it was pointed out by the Erzya evaluator that tracing back to the source of the error, many words were not translated accurately in the XML dictionaries. While polysemy causing issues was something to be expected, we cannot say for certainty how much the noise coming from the dictionaries contributes to the number of incorrect triples and how much is due to true polysemy.

Another problem pointed out by the evaluators was that the words had a wrong part-of-speech for the relation. For example, the Komi-Zyrian word *искöвтны* (time) was indicated to be an adverb by the Komi-Zyrian evaluator, and thus it cannot work as a direct object, even though the word was marked as a noun in the original dictionary.

6 Discussion and Future Work

In this paper we have presented how we have built a semantic database with syntactic information for Finnish automatically and how this database has been translated into four minority languages. The semantic databases are a first step towards the applicability of statistical methods in the context of Uralic languages that have mainly received interest in the rule-based approach to NLP.

Studying the overlap of the minority language dictionaries and SemFi, we found that more research can be done in the future in order to improve the coverage of the SemUr databases. One of them has to do with the rich derivational morphology of Finnish. Some common words deriving from another word were not recorded in SemFi, perhaps, because of the lemmatizer used to parse the original data. It is not uncommon to see this phenomenon in the Internet ParseBank data where a derivational word has been lemmatized into the word it has derived from. Also the number of frequentative verb forms in the Moksha dictionary points out the need to solve the lemmatization in a different way. In order to capture the semantics Moksha expresses in a lexicalized form, the Finnish frequentative forms should not be lemmatized back to the non-frequentative word form.

Compounding was also quite a challenge, because there is not a linguistic limit to what words can be used to form a compound, neither is there a linguistic limit to how many words can be compound together. An interesting question for the future would be, how we can predict the syntactic relations of a compound word if we know its part-of-speech and the relations of the words in the compound. More often than not, the meaning of a compound is compositional and can be derived by the meaning of its constituents. In other words, theoretically one should be able to predict the syntactic relations of an unknown compound by the relations of each individual part of it.

Structural differences between the languages are a minor source of error, but polysemy is a much bigger issue in the direct translation. This could potentially be mitigated by using multiple majority languages for the projection of the syntactic-semantic knowledge. A semantic distinction not made by Finnish, might be captured by Russian and vice versa.

One of the future directions of research is to apply SemFi and SemUr in other NLP tasks. For example, the use of these databases in improving the dictionary combination task for the same languages will be studied in the future. Another interesting possibility is to use the databases in natural language generation tasks, especially in generating parallel data for these languages. Generated parallel data together with monolingual data could be used in tasks such as neural machine translation.

7 Release of the Data

Both SemFi 2.1[2] and SemUr 1.1[3] described in this paper have been made publicly available on Zenodo under the CC BY license. The recommended way of accessing these databases is by using the functionality provided in the UralicNLP[4] Python library.

References

1. Alnajjar, K., Hadaytullah, H., Toivonen, H.: Talent, Skill and Support. A Method for Automatic Creation of Slogans. In: Proceedings of the Ninth International Conference on Computational Creativity. pp. 88–95 (2018)
2. Antonsen, L., Johnson, R., Trosterud, T., Uibo, H.: Generating Modular Grammar Exercises with Finite-state Transducers. In: Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013. pp. 27–38. No. 086, Linköping University Electronic Press (2013)
3. Fares, M., Kutuzov, A., Oepen, S., Velldal, E.: Word Vectors, Reuse, and Replicability: Towards a Community Repository of Large-text Resources. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa. pp. 271–276. No. 131, Linköping University Electronic Press (2017)
4. Hämäläinen, M.: Harnessing NLG to Create Finnish Poetry Automatically. In: Proceedings of the Ninth International Conference on Computational Creativity. pp. 9–15 (2018)
5. Hämäläinen, M., Rueter, J.: Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages. In: Proceedings of the Eighteenth EURALEX International Congress. pp. 967–978 (2018)
6. Hämäläinen, M., Rueter, J.: Development of an Open Source Natural Language Generation Tool for Finnish. In: Proceedings of the Fourth International Workshop on Computational Linguistics for Uralic Languages. pp. 51–58 (2018)

[2] https://zenodo.org/record/1463685
[3] https://zenodo.org/record/1463688
[4] https://github.com/mikahama/uralicNLP

7. Hämäläinen, M., Tarvainen, L.L., Rueter, J.: Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA) (2018)
8. Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., Ginter, F.: Building the Essential Resources for Finnish: the Turku Dependency Treebank. Language Resources and Evaluation 48(3), 493–531 (2014)
9. Hyvönen, E.: Finnonto-malli kansallisen semanttisen webin sisältöinfrastruktuurin perustaksi-visio ja sen toteutus. Kansallinen ontologiapalvelu ONKI pilottikäyttöön-julkistustilaisuus. Teknillinen korkeakoulu, Espoo 12, 2008 (2008)
10. Laippala, V., Ginter, F.: Syntactic N-gram Collection from a Large-scale Corpus of Internet Finnish. In: Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014. vol. 268, p. 184. IOS Press (2014)
11. Lindén, K., Carlson, L.: FinnWordNet-WordNet på finska via översättning. LexicoNordica (2010)
12. Moseley, C. (ed.): Atlas of the World′s Languages in Danger. UNESCO Publishing, 3rd edn. (2010), online version: http://www.unesco.org/languages-atlas/
13. Moshagen, S.N., Pirinen, T.A., Trosterud, T.: Building an Open-source Development Infrastructure for Language Technology Projects. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16. pp. 343–352. No. 85, Linköping University Electronic Press (2013)
14. Pirinen, T.A., Listenmaa, I., Johnson, R., Tyers, F.M., Kuokkala, J.: Open morphology of Finnish (2017), http://hdl.handle.net/11372/LRT-1992, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University