Semi-automatic Enhancement of Bidictionary from Aligned Sentences

Tiina Puolakainen

Institute of the Estonian Language**
tiina.puolakainen@ut.ee

Abstract. When the base of rule-based machine translation system is established and it already can give some reasonable translations - at some point of development of rule-based MT system the insufficiency of the dictionary becomes a bottleneck, as available digital lexical resources are already utilized. The issue holds especially considering low-resource and morphologically rich languages with very productive compounding formation.
The article proposes a method for generating new bidictionary entries for closely related morphologically rich languages with productive compounding. The idea is to use incompletely translated sentences along with their correct translations to incrementally improve and enhance the translation dictionary. The method employs also the advantages of rule-based machine translation taking into account the inner bitranslation state to find the interconnections between source and target utterances, generate new entries and minimize the unmatched existing entries of bidictionary with mono-dictionaries.

1 Introduction

The aim of the proposed method is to compare target language and MT translated sentences to identify missing or mistranslated words or multi-word expessions and generate missing bilingual dictionary entries. For morphologically rich languages the task becomes more complicated due to the additional subtask of finding correct lemmas for the dictionary entries and in case of compounds or multi-word expressions also the exact morphological description of unchangable parts. The lemmas and morphological description of generated entries should be chosen in such a way that they would fit into concrete system, e.g. comply with the previous source language analysis, the next transfer steps and target language generation stage and are properly handled by bilingual mapping which imposes additional conditions on the process. Additionally, by adding new translations, that could contain even quite unexpected correspondences from specific contexts, the urgent need for lexical selection component emerges which could make reasonable choice of available translations in particular context.

** The main work was conducted being empoyed by the University of Tartu and The Arctic University of Norway

For every system consisting of multiple chained processes, decisions made earlier influence all the next stages and so every error is escalating other errors on the next stages. So the quality of rule-based machine translation system depends on each of the stages and the aim is to generate new entries introducing as few errors as possible. Therefore the proposed method tries to use as much information as is available provided by the rule-based machine translation components and at the same time is quite self-contained and do not rely on external sources except some adjustable parameters.

The rest of the article is organised as follows: the next chapter introduces the stucture of the rule-based machine translation (RBMT) system where the proposed method is designed to work and the 3rd chapter describes the method of generation new lexical entries and applying them during translation process. The following chapter gives an overview of evaluation of the method on the particular Finnish-Estonian MT system prototype and last chapter concludes.

2 RBMT Environment

The method is designed bearing in mind to be used in the rule-based machine translation system, in particular the Apertium[1] free/open-source RBMT platform [1] for developing transfer-based machine translation systems. For the current state-of-the-art 43 language pairs have released translator versions and are considered to be stable[2] and many more are in the development. The translation for the language pair in the Apertium system goes through multiple stages:

- deformatting and tokenizing the source language input text,
- morphological analysis of the source language tokens by means of the Finite State Transducers [2] (fst),
- morphological disambiguation, in particular case by means of Constraint Grammar disambiguator [3],
- pretransfer and lexical transfer based on bilingual dictionary,
- default lexical selection component,
- multiple possible steps of applying structural transfer rules,
- generation of target language surface forms from morphological description,
- formatting the target language output accordingly to original format.

3 The Method Description

For finding the relations between the source, target language and MT utterances the following sources are used: morphologically analysed source language, morphologically analysed target language human translation, MT intermedate bilingual translation file with possible translations for source language words, target language MT translation with morphological analyses for form generation and MT translation surface forms. All the inputs contain the information

[1] https://www.apertium.org
[2] http://wiki.apertium.org/wiki/

fragments that taken together enable to compare MT and intended translations and generate missing bilingual dictionary entries. Due to the differences in morphological analysis, derivational and compounding issues none of connections between sources are fully deterministic, but the most problematic is to relate MT translation with corresponding intended human translation sentence words.

The analysis and generation parts of the method consist of:

- (1) Harmonization of inputs which are manually aligned sentences in 3 different formats: plain texts, morphologically analysed texts and a special bilingual intermediate translation form. For more convenient use they are all transformed to fit into unified structure.
- (2) Chunking – relies on the 'simplified' input supposing that target sentences are precise human translations. At this stage assume that the intrasentential punctuation is almost the same and is used to delimit sentence to smaller parts. If the assumption does not hold, then other algorithms or, alternatively, entire sentences should be used.
- (3) Finding relations between words in parallel phrases – the most important part, the quality of found relations will imply also the quality of generated dictionary entries. The method takes an advantage of the simplified human cognitive approach of reading the text in foreign language: first identify the "known" words and then try to fill the "gaps". In the current MT context: when relations with high confidence are found between MT translated and target human translated aligned sentences' words, then try to build the relations between remaining words starting from using more reliable rules to some heuristics, for example, the continuation principle, which causes to prefer longer connected sequences of words. The morphological information and some fixed patterns are used as more reliable rules. Coordinating conjunctions are used at this step as additional clue for splitting long phrases, but this relies on the knowledge that particular languages in the pair are closely related. The connecting of the compound parts in two languages is somewhat tricky as compound on one side can be or not to be a compound on the other side and if it is not a compound, it can still correspond to one or more separate words. It can be also some fusion as a three-parts compound on one side and compound plus one separate word on the other side. In addition, due to morphological analysis and disambiguation it can be represented as one or more words or word parts.
- (4) Generation of bilingual dictionary entries if correct translations are not found in the bilingual dictionary. Only the lemmas, that correspond to source language analyser output and are accepted by target language generator, should be used on both source and target sides of translation entry correspondingly: if the translation word is not contained in the target language fst, then the correct form would not be generated. For the compounds or multiword expressions also the morphological description for unchangeable parts need to be provided and homonym's lemmas should contain the homonym identifier.
- (5) Extracting context information for source language with multiple possible translations. Weight of closer context is higher as well as lexical match

has also much higher weight than the morphological form, although in some cases of specific constructions the morphology becomes more important as, for example, the Finnish verb *pitää*, which has many different meanings (most usual 'must', 'have to', also 'hold' or 'have'), in a specific morphological context translates to Estonian as *meeldima* ('to like'). For these purposes context conditions for the rules can be edited by a linguist. The very frequent but functional words can be ignored as a relevant context. Only source language is used as a context for selecting translation for word in question, because it is more reliable as represents original utterance while translated content is the object of change and so is much more variable and besides that can also contain some morphological disambiguation or other errors.

The application stage for the found new dictionary entries includes additional steps after pretransfer and bilingual translation mapping for splitting compounds, because compounds with multiple inflected parts are not supported by standard procedure, and selecting appropriate translations from bidictionary alternatives based on extracted context.

4 Evaluation

For the evaluation the proposed method was applied in the setting of the rule-based Finnish-Estonian machine translation (apertium) prototype, the morphological analysis and synthesis of languages are provided by Giellatekno[4][3] language resources[4]. These two closely related languages belong to the Finno-Ugric language family and both have a morphologically rich structure (to mention 14 cases for nominals and over 30 forms in verb paradigm) with plentiful use of derivational and compounding mechanisms and relatively free word order, but only a small very core commonly comprehensible lexical base, with lots of false friends. To illustrate the degree of freedom of the word order we will check the assumption[5] that out of all 24 possible complete matchings between two constituents of length four each it was not possible to find real examples of constituent arguments undergoing "inside-out" transpositions: for example, the translation of *As the evening was boring, they went to the theatre yesterday* to Estonian *Kuna õhtu oli igav, läksid nad eile teatrisse* contains such "inside-out" transposition in the subordinated clause with word correspondences *they* $\quad$ *nad*, *went* $\quad$ *läksid*, *to the theatre* $\quad$ *teatrisse*, *yesterday* $\quad$ *eile*.

As a sort of adaptation for this language pair it was decided to ignore some sets of words when relating MT and target translation sentence words with each other with the following considerations:

– personal pronouns – 1) a closed set, so not a focus for lexical expansion, although they can be translated for another pronoun for example; 2) widely used posessive suffixes in Finnish are usually translated to personal pronouns,

[3] http://giellatekno.uit.no/
[4] https://victorio.uit.no/langtech/trunk/langs/fin and /langs/est

that makes their use more asymmetric in two languages and adds complexity for establishing relations,
- auxiliary verb *olla* (est *olema*, 'to be') – not important for the lexical task,
- adpositions (pre- or postpositions) – the set of adpositions is not completely closed, it can vary in the context if it is considered to be an adposition or a noun, also as a disambiguation mistake. But there are differences in using adpositions in two languages and as functional words they are also not in the focus for the task (although can be used for relation decisions).

It appeared also useful to introduce part-of-speech relaxed correspondence – when comparing part-of-speech for decision to relate two words with each other or not, some relaxation rules can be drawn, for example the set of common noun, proper noun and acronym is considered to be interchangeable as it depends a lot on the choices of morphological analyser, to which extent it distinguishes between them and also on disambiguation choices in particular context. For similar reasons the equivalence sets for some morphological forms are introduced. Although case system is very similar in two languages, the usage of particular cases is not always identical. But one have to be very careful with definition of such equivalence classes and use them only in very systematic cases, not for all occasional collisions or mistakes.

4.1 Evaluation results on the development set

The development set consisted of 2045 parallel human translated sentences from textbook for learning Finnish, for every sentence only one human translation variant was available. The approximate number of words is 12550 for development and 1050 for test set – the numbers are in MT translation 'atomic' words where the parts of compounds are counted separately if they could be distinguished by the analysis, so even the number of all words can vary a bit for different system runs depending on the translation result.

For evaluation purposes every word was assigned with one of the following categories:

- correct lexical translation,
- lexical selection error – the correct choice was available but not selected,
- incorrect lexical translation – do not include lexical selection errors,
- ignored – words from predefined ignore-sets or with ignore-POS,
- not related – words that were not related with any target translation word, because the relations tried to be made with higher confidence, so in unclear circumstances the relation is abandoned.

The lexical selection and incorrect lexical translation errors are distinguished because they need different kind of action – in the first case it is the question of right choice from possible translations and in the other the new lexical translation have to be generated and in case it becomes not the only one, then also the context for lexical selection extracted. There is one more reason for viewing lexical selection errors separately – in case of synonyms it may be not an error at

all but a question of subjective preference, but this question is not investigated further in current task because of belief that it is not trivial to draw the line between synonyms and different (sub)meanings considering very different specific contexts. Only lexical identity is important for this task, so the correctness of the morphological form is not considered – word with correct lemma but in wrong case was assigned the 'correct' category. For the same reason also the opposite case was also considered 'correct' – if the lemma was different but surface form exactly the same, because of possible different possible morphological analysis possibilities. For example, the participles can be viewed as verb forms and therefore have a verb lemma with verb reading morphological tags or an adjective with surface form lemma with its own tags, and, after all, it is exactly the wordform and for human reading it makes exactly the same effect.

For the developmet set, applying the enhanced dictionary with derived lexical selection rules decreases the number of the incorrectly translated words drastically from 3196 to 627 (by 2569, for 20.6% of all words) and increases the number of corrects from 6793 to 9117 (by 2324, for 18.6% of all words), as can be seen in Table 1. The number of lexical selection errors grows twice (for 3.6% of all words), but this is much smaller number compared to the advantages in changes of corrects and incorrects and this is more 'soft' error than totally wrong or unrecognized word.

Table 1. Results of applying enhanced dictionary on the development data set.

Dictionary	**Correct**	**(Lexsel+)**	**Lexsel**	**Lexsel-**	**Incorr**	**Ignored**	**Notrel**	**All**
Default	6793	0	465	0	3196	1761	200	12415
	54.7%	0%	3.7%	0%	25.7%	14.2%	1.6%	100%
Develop.	9117	5099	52	857	627	1780	145	12578
	72.5%	40.5%	0.4%	6.8%	5.0%	14.1%	1.1%	100%

The lexical selection errors are divided into two columns: Lexsel column shows the number of errors where the choice was made by the default system taking the first available translation from the set at the moment of translation and Lexsel- column shows the number of errors where the choice was made using introduced lexical selection by the extracted context. As it is seen, this is not perfect yet and needs more improvement. The number of correctly applied lexical selections by the new extracted context component is shown in (Lexsel+) column and it has a fundamental impact on the increase of corrects. For example, the translation of Finnish sentence *Hän soittaa pianoa* with default dictionary was *Ta helistab klaverit*, which is expected to mean 'He/She calls (by phone) the piano'. With the new lexical selection component exploiting automatically extracted context it receives the correct translation *Ta mängib klaverit*, that means 'He/She plays piano'. The numbers in this column (Lexsel+) are included also in the numbers of Corrects, so they do not add separately to the sum of All.

4.2 Evaluation results on the test set

The test set of 120 parallel human translated sentences from Finnish Turku Dependency Treebank[6] (TDT). Evaluation results show gradually improve for corrects, lexsels and incorrects from applying default dictionary state through development set enhanced dictionary to further enhanced dictionary from the test set, with minor changes in ignored and not-related words, as it is shown in Table 2. The number of correctly translated words grows from 451 to 697 (for 23.6% of all words) and incorrects number decreases from 369 to 74 (for 28.3% of all words). The first row shows results for default dictionary state for the moment of the test, the second row gives results for dictionary and extracted context enhanced by the development set and, finally, the third row contains the results of improved dictionary and extracted context enhanced by the test set (additionally to the development set). The new lexical selection component adds 133 errors in the second round while giving also 245 correct suggestions; in the third round the number of the lexical selection errors is decreasing to 110 with the profit of 482 correct selections. The unexpected grow of traditional lexical selection errors to 9 is due to 4 intances of *ettepanek* 'proposal', being formally analysed differently as a whole word or a compound *ette+panek*, while retaining the sense as a fixed compound.

Table 2. Results of applying enhanced dictionary on the test data set.

Dictionary	**Correct**	**(Lexsel+)**	**Lexsel**	**Lexsel-**	**Incorr**	**Ignored**	**Notrel**	**All**
Default	451	0	24	0	369	133	51	1028
	43.9%	0%	2.3%	0%	35.9%	12.9%	5.0%	100%
Develop.	478	245	7	133	239	135	40	1032
	46.3%	23.7%	0.7%	12.9%	23.2%	13.1%	3.9%	100%
Dev&test	697	482	9	110	74	140	30	1055
	66.1%	45.7%	0.9%	10.4%	7.0%	13.3%	2.8%	100%

For the sake of the ensuring backbone of the method functionality the test was made to apply translation system with the test-data-enhanced dictionary and test-data-enhanced extracted context to the development set to see what kind of effect it will make. As a result even with as small test-data as it was of 120 sentences (used to derive the entries additionally to development set), the initial results improved by 29 in corrects and incorrects decreased by 123 words, as shows Table 3.

As show the test results the method is able to successfully identify relations between source and target sentences, derive new dictionary entries with correct lemmas and part-of-speeches, enabling to include them into analysis-transfer-generation translation process, including also some treatment of the compounds. The procedure is safe in the sense that applying further enhanced dictionary and

Table 3. Results of applying newly enhanced dictionary on the initial development data set.

Dictionary	Correct	(Lexsel+)	Lexsel	Lexsel-	Incorr	Ignored	Notrel	All
Develop.	9117	5099	52	857	627	1780	145	12578
	72.5%	40.5%	0.4%	6.8%	5.0%	14.1%	1.1%	100%
Dev&test	9146	5558	50	956	504	1780	142	12578
	72.7%	44.2%	0.4%	7.6%	4.0%	14.2%	1.1%	100%

extracted context on the initial development data showed the improvement of the results.

4.3 Double-check evaluation results

To ensure that there is no bias with particular development set in the evaluation results, an additional test was made with different development set, consisting of 1663 sentences from Finnish Turku Dependency Treebank (TDT). Another aim was to see if there is a difference in using slightly bigger set of more real sentences for development. The derived dictionary and lexical selection context rules were applied to the same test set as in the previous experiment and analogously the test set enhanced dictionary was checked on the initial development set. The evaluation results are presented in Table 4 and Table 5 accordingly.

Table 4. Results of applying enhanced dictionary on the TDT development data set.

Dictionary	Correct	(Lexsel+)	Lexsel	Lexsel-	Incorr	Ignored	Notrel	All
Default	7285	0	504	0	6263	1836	556	16444
	44.3%	0%	3.1%	0%	38.1%	11.2%	3.4%	100%
Develop.	11423	7256	140	1955	1210	1872	380	16980
	67.3%	42.7%	0.8%	11.5%	7.1%	11.0%	2.2%	100%
Dev&test	11420	7603	141	2033	1136	1869	379	16978
	67.3%	44.8%	0.8%	12.0%	6.7%	11.0%	2.2%	100%

Seemingly the results are slightly poorer comparing to the first textbook development set (Table 1 and Table 3), but considering that for textbook the baseline contained already 54.7% corrects and only 25.7% incorrects while the baseline for TDT development set consists of 44.3% corrects and 38.1% incorrects, it becomes clear that the improvement in the latter case is better – by 23% of corrects istead of 18% and by 19% instead of 14%, considering incorrects and new lexical selection component together.

The comparison with the test results of the first experiment (Table 2) show very close change for corrects (with deviation in 1.5%), but more substantial

Table 5. Results of applying TDT enhanced dictionary on the test data set.

Dictionary	Correct	(Lexsel+)	Lexsel	Lexsel-	Incorr	Ignored	Notrel	All
Default	451	0	24	0	369	133	51	1028
	43.9%	0%	2.3%	0%	35.9%	12.9%	5.0%	100%
Develop.	496	275	3	183	179	136	42	1039
	47.7%	26.5%	0.3%	17.6%	17.2%	13.1%	4.0%	100%
Dev&test	689	505	8	122	66	137	33	1055
	65.3%	47.9%	0.8%	11.6%	6.3%	13.0%	3.1%	100%

difference for errors: the number of errors for development dictionary nad context rules is approximately the same – around 35–36%, but if in the first case it consists of 23% of lexical errors and 13% of lexical selection mistakes, then in the latter case these components are about the same 17%, which means the less number of strict errors.

4.4 Cross-validation of different development paths

Additional tests were performed to check how differently induced dictionaries will behave on a bigger texts, for better comparison the enhanced dictionaries were applied mutually on the other development set and further extended dictionaries on the test set. The results of the experiment are shown in Table 6.

The first row shows the application of the first textbook induced dictionary on the TDT development set and second row with further enhanced dictionary on the same TDT set, the third and fourth rows accordingly the application of TDT induced and enhanced dictionaries on the textbook set. Both directions the initial result improves over the baseline, the TDT derived dictionary and context set performs a bit better due to larger corpus with more complicated sentences (comparing rows 1 and 3 with data in Table 4 and Table 1 respectively) and achieves comparable results with enhancing from development set itself (comparing rows 2 and 4 with data in Table 4 and Table 3 respectively).

The fifth and sixth rows contain the data of application of dictionaries induced from both development sets, one starting from textbook and the other from TDT development set. The last seventh row contains application of dictionary that was induced from TDT, enhanced from textbook and further on from test set itself. Comparing with the previous results in Table 2 and Table 5 shows that in both cases the dictionaries and context rules from double development set outperformed the first experiments, as it could be expected. Deriving additional information from self does not give as clear advantage showing that in previous cases it could give some overfitting effect.

4.5 Combinatorics of compounds

There are multiple clues for identifying and relating the parts of compounds in the different 'views' as well as decision on generating new or additional bilingual

Table 6. Results of cross-validation on the development and test data sets.

Dictionary	Correct (Lexsel+)	Lexsel	Lexsel-	Incorr	Ignored	Notrel	All
Txtbook	7681	3888	136	2543	3858	1834	504	16556
	46.4%	23.5%	0.8%	15.4%	23.3%	11.1%	3.0%	100%
Txb&enh.	11290	7910	143	2129	1131	1912	366	16971
	66.5%	46.6%	0.8%	12.5%	6.7%	11.3%	2.2%	100%
TDT dev.	6917	3624	79	1684	1817	1761	168	12426
	55.7%	29.2%	0.6%	13.6%	14.6%	14.1%	1.3%	100%
TDT&enh.	8959	5988	59	1079	597	1763	126	12583
	71.2%	47.6%	0.5%	8.6%	4.7%	14.0%	1.0%	100%
Test Txb	510	340	3	203	152	136	37	1041
	49.0%	32.7%	0.3%	19.5%	14.6%	13.0%	3.5%	100%
Test TDT	507	345	4	203	152	139	36	1041
	48.7%	33.1%	0.4%	19.5%	14.6%	13.3%	3.5%	100%
Test TDT+	684	531	10	129	64	142	29	1058
	64.7%	50.2%	0.9%	12.2%	6.0%	13.4%	2.7	100%

dictionary entry. The compounding processes in the both languages under discussion are very productive and often compound on one side is also translated as xompound on the other side, but not always. Usually it is possible to translate parts of the compound independently, but often the compound restricts the translation from multiple possibilities of single word translations to the specific one. Of course, there are special cases where such tactic does not work and even if parts have meaningful translations in other language, the actual translation consists of totally different words, those translations should not be included in the dictionary as single word translations as they would be wrong. For example, Finnish *päiväkoti* (eng kindergarten) translates to Estonian as *lasteaed* (composition of "children's garden"), but translating the parts of the compound directly it gives "day's home", which for Estonian would also give an existing compound *päevakodu*, but it means a very different thing (as well as would mean *lastekodu* "children's home").

The most common case is if compound has two parts in both languages and it is analysed as two in target and (bi)translation notations, but is analysed as one word in source language analysed notation. The reason for that is in the principle that Finnish finite state transducer has most of compounds lexicalized but the Estonian finite state transducer has mostly dynamic compound formation. For that reason there is no sense to add translations from whole to whole word as in target language generator this lemma would be missing and the correct form would not be generated. So the translation should translate compound by its parts unless the target language compound is lexicalized and contained in monolingual dictionary.

It is possible, that in source language compound will be unrecognized due to intersection of source finite state analyser transducer with bilingual dictionary

transducer, if bilingual dictionary do not contain the compound not as a whole nor by parts. In this case connections should be found through the separate morphologically analyzed source that uses the automata without intersection with bilingual dictionary.

Finally, it is possible that the translation of compound as a whole to whole word is contained in a dictionary already, but if it is not included as a one word into the target language fst generator, then the correct forms cannot be generated and so another translation have to be added into the bilingual dictionary that consists of words existing in target language fst generator. Usually non-final parts of compound are inflected or have special prefix-form and therefore the morphological description should be added into the dictionary entry for each compound part separately.

4.6 Derivatives

Derivative forms are also very common in both languages and some of derivational patterns are very similar: present and past tense participles for active and passive voice (traditionally called personal and impersonal voice in Estonian), *-sti* adverb formation in Finnish and corresponging *-lt* in Estonian (*-lly* in English), *-minen* noun formation from verb in Finnish and *-mine* in Estonian (*-tion* in English).

For smooth translation some decisions are to be made on the way the analysis is presented in morphological finite state transducers. For some forms as participles the general solution is to give the verb lemma also as lemma of participle with corresponding analysis and that works very well. For other forms general decisions are not so straightforward. From the one hand, it is good to have the initial lemma and explcit derivation path in the anaysis not to multiply the number of entries of morphological and bilingual dictionaries. But this presupposes that analogical derivation process exists also in target language and can be successfully mapped. From the other hand, such approach can create some unintended mess for human readability, morphological disambiguation and translation due to different lemmas and parts of speech instead of usual ones, especially if other language don't use derivational patterns in the same way. The right balance needs to be found.

5 Conclusions

Considering the generation of the translation dictionary entries the purpose can be of two kinds: to generate as much as possible with more likely errors to filter out afterwards or try to avoid errors generating correspondences still with possible occasional mistakes. The current approach aims for the latter and is resource consuming, taking into account as much information from previous processing steps as possible to connect source with MT and target human translation, reveal lexical translation inconsistencies and generate new bilingual dictionary translation entries along with the extraction of contexts for correct lexical selection

where multiple choices become available. Along with RBMT system's intermediate bilingual translation state the current approach is using the morphologically analysed and disambiguated source, target language and MT translated sentences and showes promising results, being able to identify and extract relations with correct lemmas and exact morphological descriptions for dictionary entries, including compounds, although leaving relating decisions undone in unclear contexts to possibly avoid errors. Therefore the method can be used for supervised self-learning, applied iteratively to incrementally improve bilingual dictionary. The further advancements could involve at least the syntactic analyses of sources if not use of some semantic classifiers, better treatment of synonyms, compounds and multi-word expressions along with lexical selection from multiple possible translations.

Acknowledgments

This work has been supported by Estonian Ministry of Education and Research (grant IUT 20-56 "Eesti keele arvutimudelid (Computational Models for Estonian)" and Norwegian-Estonian Research Cooperation Programme (grant EMP160 "SAMEST – Sami-Estonian language technology cooperation – similar languages, same technologies").

References

1. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M.: Apertium: a free/open-source platform for rule-based machine translation. Machine translation **25** (2011) 127–144
2. Koskenniemi, K.: Two-level morphology—A General Computational Model for Word-Form Recognition and Production. PhD thesis, Department of General Linguistics. University of Helsinki, Finland (1983)
3. Karlsson, F., Voutilainen, A., Heikkilae, J., Anttila, A.: Constraint Grammar: a language-independent system for parsing unrestricted text. Volume 4. Walter de Gruyter (1995)
4. Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T., Tyers, F.M.: Open-source infrastructures for collaborative work on under-resourced languages. In: LREC 2014, Workshop: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, Leuven, Belgium (2014) 71–77
5. Wu, D.: Grammarless extraction of phrasal translation examples from parallel texts. In: Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven, Belgium (1995) 354–372
6. 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 (2013) 1–39