Morphological Segmentation and OPUS for Finnish-English Machine Translation

Jörg Tiedemann[1], **Filip Ginter**[2], **and Jenna Kanerva**[2,3]
[1] Department of Linguistics and Philology, Uppsala University, Sweden
[2] Department of IT, University of Turku, Finland
[3] University of Turku Graduate School (UTUGS), Finland
jorg.tiedemann@lingfil.uu.se, figint@utu.fi, jmnybl@utu.fi

Abstract

This paper describes baseline systems for Finnish-English and English-Finnish machine translation using standard phrase-based and factored models including morphological features. We experiment with compound splitting and morphological segmentation and study the effect of adding noisy out-of-domain data to the parallel and the monolingual training data. Our results stress the importance of training data and demonstrate the effectiveness of morphological pre-processing of Finnish.

1 Introduction

The basic goal of our submissions is to establish some straightforward baselines for the translation between Finnish and English using standard technology such as phrase-based and factored statistical machine translation, in preparation for a more focused future effort in combination with the state-of-the-art techniques in SMT for morphologically complex languages (see e.g. (Fraser et al., 2012)). The translation between Finnish and English (in both directions) is a new task in this year's workshop adding a new exciting challenge to the established setup. The main difficulty in this task is to manage the rich morphology of Finnish which has several implications on training and expected results with standard SMT models (see the illustration in Figure 1). Moreover, the monolingual and parallel training data is substantially smaller which makes the task even tougher compared with other languages pairs in the competition. In our contribution, we focus on Finnish-English emphasizing the need of additional training data and the necessity of morphological pre-processing. In particular, we explore the use of factored models with multiple translation paths and the use of morphological segmentation based on proper morphological annotation and simple rule-based heuristics.

Syksyllä taidemuseossa avataan uudet näyttelyt
Autumn+ADE art_museum+INE open+PASS new+PL exhibition+PL
In_autumn in_art_museum will_be_opened new exhibitions
New exhibitions will be opened in the art museum in autumn

Figure 1: A sentence illustrating the inflective and compounding nature of Finnish in contrast to English. (ADE, INE: adessive, inessive cases, PASS: passive, PL: plural)

We also add noisy out-of-domain data for better coverage and show the impact of that kind of data on translation performance. We also add a system for English-Finnish but without special treatment of Finnish morphology. In this translation direction we only consider the increase of training data which results in significant improvements without any language-specific optimization.

In the following, we will first present our systems and the results achieved with our models before discussing the translation produced in more detail. The latter analyses pinpoint issues and problems that provide valuable insights for future development.

2 Basic Setup and Data Sets

All our translation systems are based on Moses (Koehn et al., 2007) and standard components for training and tuning the models. We apply KenLM for language modeling (Heafield et al., 2013), fast_align for word alignment (Dyer et al., 2013) and MERT for parameter tuning (Och, 2003). All our models use lowercased training data and the results that we report refer to lowercased output of our models. All language models are of order five and use the standard modified Kneser-Ney smoothing implemented in KenLM. All phrase tables are pruned based on significance testing (Johnson et al., 2007) and reducing translation options to at most 30 per phrase type. The maximum phrase length is seven.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 177–183,
Lisboa, Portugal, 17-18 September 2015.

For processing Finnish, we use the Finnish dependency parser pipeline[1] developed at the University of Turku (Haverinen et al., 2014). This pipeline integrates all pre-processing steps that are necessary for data-driven dependency parsing including tokenization, morphological analyses and part-of-speech tagging, and produces dependency analyses in a minor variant of the Stanford Dependencies scheme (de Marneffe et al., 2014). Especially useful for our purposes is the morphological component which is based on OMorfi - an open-source finite-state toolkit with a large-coverage morphology for modern Finnish (Lindén et al., 2009). The parser has recently been evaluated to have LAS (labeled attachment score) of 80.1% and morphological tagging accuracy of 93.4% (Pyysalo et al., 2015).

The data sets we apply are on the one hand the official data sets provided by WMT and, on the other hand, additional parallel corpora from OPUS and large monolingual data sets for Finnish coming from various sources. OPUS includes a variety of parallel corpora coming from different domains and we include all sources that involve Finnish and English (Tiedemann, 2012). The most important corpora in terms of size are the collection of translated movie subtitles (OpenSubtitles) and EU publications (DGT, EUbookshop, EMEA). Some smaller corpora provide additional parallel data with varying quality. Table 1 lists some basic statistics of Finnish-English corpora included in OPUS. The final two rows in the table compare the overall size after cleaning the corpora with the pre-processing scripts provided by Moses with the training data provided by WMT for Finnish-English. We can see that OPUS adds a substantial amount of parallel training data, more than ten times as many sentence pairs with over six times more tokens. A clear drawback of the data sets in OPUS is that they come from distant domains such as movie subtitles and that their quality is not always very high. User contributed subtitle translations, for example, include many spelling errors and the alignment is also quite noisy. EUbookshop and EMEA documents are converted from PDF leading to various problems as well (Tiedemann, 2014; Skadiņš et al., 2014). Software localization data (GNOME, KDE4) contains variables and code snippets which are not appropriate for the WMT test domain. One of the main questions we wanted to answer with our experiments is whether this kind of data is useful at all despite the noise it adds.

corpus	sentences	en-words	fi-words
Books	3.6K	69.7K	54.5K
DGT	3.1M	61.8M	46.9M
ECB	157.6K	4.5M	3.4M
EMEA	1.1M	14.2M	11.9M
EUbookshop	2.0M	51.4M	37.6M
JRC-Acquis	19.7k	388.7k	273.6k
GNOME	62.2K	313.3K	254.6K
KDE4	108.1K	596.0K	578.6K
OpenSubtitles	110.1K	856.3K	604.7K
OpenSubtitles2012	12.9M	111.5M	74.4M
OpenSubtitles2013	9.8M	87.8M	55.7M
Tatoeba	12.2K	103.2K	77.0K
WMT-clean	2.1M	52.4M	37.6M
OPUS-clean	29.4M	328.1M	227.6M

Table 1: Finnish-English data in OPUS. WMT-clean and OPUS-clean refer to the entire parallel training data set from WMT and OPUS, respectively, after pre-processing with the standard Moses cleanup script.

Table 1 also illustrates the morphological differences between English and Finnish. Based on the token counts we can clearly see that word formation is quite different in both languages which has significant implications for word alignment and translation. Due to the rich morphology in Finnish we expect that adding more training data is even more crucial than for morphologically less complex languages. To verify this assumption we also include additional monolingual data for language modeling for the English-Finnish translation direction taken from the Finnish Internet Parsebank,[2] a 3.7B token corpus gathered from an Internet crawl and parsed with the abovementioned dependency parser pipeline (Kanerva et al., 2014). For English we include the fifth edition of the LDC Giga-Word corpus.

3 Factored Models for Finnish-to-English

Our baseline models apply a standard pipeline to extract phrase-based translation models from raw lowercased text. We use constrained settings with WMT data only and unconstrained settings with additional OPUS data. Our primary systems apply factored models that include three competing translation paths:

- Surface form translation

[1] http://turkunlp.github.io/Finnish-dep-parser

[2] http://bionlp.utu.fi/finnish-internet-parsebank.html

- Translation of lemmatized input
- Translation of lemmatized and morphosyntactically tagged input

The unconstrained system replaces the first translation path with a phrase table extracted from the entire corpus including all OPUS data. However, we did not parse the OPUS data and take the other two models from WMT data only. We tuned our systems with half of the provided development data (using every second sentence) and tested our models on the other half of the development data. Table 2 lists various models that we tested during development and the various components are explained in more detail in the sections below.

system	*BLEU*
constrained	
baseline	16.2
factored	17.8
factored+pseudo	18.2
unconstrained	
baseline+WordNetTrans	16.5
baseline+WordNetTrans&Syn	16.6
baseline+opus	19.0
baseline+opus+WordNetTrans	19.1
baseline+opus+WordNetTrans&Syn	19.1
factored+opus	19.2
factored+opus+pseudo	19.9
factored+opus+pseudo+word2vec	20.0
factored+opus+pseudo+WordNetSyn	20.1

Table 2: The performance of various Finnish-English translation models on development data. *Pseudo* indicates the use of inflection pseudo-tokens, *word2vec* refers to the use of word2vec synonyms and *WordNetSyn* refers to the inclusion of WordNet synonyms for out-of-vocabulary words. *WordNetTrans* refers to translations added from the bilingual Finnish-English WordNet for OOV words.

3.1 Inflection Pseudo-Tokens

Due to the highly inflective nature of the language, a Finnish morphological marker often corresponds to a separate English word. This is especially prominent for many Finnish cases which typically correspond to English prepositions. For example, the Finnish word *talossakin* has the English translation *also in a/the house* where the inessive case (*ssa* marker) corresponds to the English preposition *in* and the clitic *kin* corresponds to the English adverb *also*. To account for this phenomenon, we pre-process the Finnish data by inserting dummy tokens for certain morphological markers, allowing them to be aligned with the English words in system training phase. These dummy tokens are always inserted in front of the text span dominated by the word from which the token was generated in the dependency parse. Thus, for instance, the case marker of the head noun of a nominal phrase produces a dummy token in front of this phrase, where the corresponding English preposition would be expected. The pseudo-tokens are generated rather conservatively in these three situations:

- a case marker other than nominative, partitive, and genitive on a head of a nominal phrase (*nommod* and *nommod-own* dependency relations in the SD scheme version produced by the parser)
- a possessive marker (eng. *my, our, etc.*) in any context
- the clitic *kin/kaan* (eng. *also*) in any context

To shed some further light on the effectiveness of the pseudo-token generation, we carry out a focused manual evaluation on the test dataset. In randomly selected 100 sentences, we marked every nominal phrase head inflected in other than nominative, partitive, and genitive case and checked in the system output whether this exact phrase head was translated correctly (as judged by the annotator, not the reference translation), regardless of the correctness of the remainder of the sentence. We compare the final system with and without the dummy token generation component, in a randomized fashion such that it was not possible to distinguish during the annotation which of the two systems the translation originated from. In total, the 100 sentences contained 148 inflected phrase heads of interest. Of these, the system with pseudo-token generation translated correctly 100/148 (68%) and without pseudo-token generation 89/148 (60%). This difference is, however, not statistically significant at p=0.12 (two-tail McNemar's test). In addition to this manual evaluation, we have also observed a small advantage for the pseudo-token generation in terms of development set BLEU score. Somewhat surprisingly, we find that only 85/148 (57%) of these inflected heads were translated using a prepositional phrase in the reference translation, showing that the correspondence of Finnish cases with English prepositions is not as strong as might intuitively seem. Of those inflected heads which were translated as a prepositional phrase in the reference, 57/85 (67%) were correct for the system with pseudo-tokens and 49/85 (58%) for the system without, whereas for those that have not been

translated as a prepositional phrase in the reference, the proportions are 43/63 (68%) and 40/63 (63%). Due to the small sample size, it is difficult to draw solid conclusions but the numbers at least hint at the intuitive expectation that the pseudo-token generation would give better results especially in cases where the translation corresponds to a prepositional phrase. The overall quality of translation of inflected nominal phrase heads however leaves much room for improvement.

3.2 Compounds

Finnish is a compounding language, once again leading to a situation whereby a single Finnish word corresponds to multiple English words. Further, compounding in Finnish is highly productive and reliable translations cannot be learned but for the most common compounds. In most cases, the compounds are correctly analyzed by the Finnish parsing pipeline, including the boundaries of the lemmas which form the compound. To assist the alignment as well as the translation process itself, we split the compound lemmas into the constituent parts as a pre-processing step in the Finnish-English direction. The following example illustrates this process ("EU support for enterprises") taken from the development data:

compound:	EU-yritystukien
segmented lemma:	EU|yritys|tuki
PoS:	N
morphology:	NUM_Pl|CASE_Gen
factored segments:	EU|EU|_|_
	yritys|yritys|_|_
	tukien|tuki|N|NUM_Pl+CASE_Gen

As shown above, PoS and morphology are only attached to the final component of the compound and string matching heuristics are used to split surface forms as well based on the segmentation of the lemma.

3.3 Synonyms and Lexical Resources

One of the major problems for statistical machine translation with limited resources is the treatment of out-of-vocabulary (OOV) words. This problem is even more severe with morphologically rich languages such as Finnish. Table 3 shows the OOV ratio in the development data that we used for testing our models. We can see that the factored models significantly reduce the amount of unknown word type and tokens.

In our final setup we tried to address the problem of remaining OOVs by expanding the input with synonyms from external resources. We looked at two possible sources: distributional models trained on large monolingual data sets and manually created lexico-semantic databases. For the former, we trained distributed continuous-vector space models using the popular word2vec toolkit[3] (Mikolov et al., 2013) on the 3.7B tokens of the Finnish Internet Parsebank data, using the default settings and the skip-gram model. We tested the use of the ten most similar words for each unknown word coming from our word2vec model (according to cosine similarity in their vector representations) to replace OOV words in the input. The second alternative uses the Finnish WordNet[4] (Niemi et al., 2012) to replace OOV words with synonyms that are provided by the database. We apply the HFST-based thesaurus for efficient WordNet lookup that enables the lookup and generation of inflected synonyms.[5] Table 4 shows the statistics of unknown words that can be expanded in the development test data. The table shows that word2vec expansion has a better coverage than WordNet but both resources propose a large number of synonyms that are not included the phrase table and, hence, cannot be used to improve the translations. However, both strategies produce a large number of spurious (context-independent) synonyms and discarding them due to the lack of phrase table coverage is not necessarily a bad thing. The results of applying our two OOV-handling strategies on the same data set are shown in Table 2.

OOVs	types	tokens
constrained		
baseline	2,451 (28.7%)	2,869 (14.5%)
factored	847 (14.5%)	958 (6.7%)
unconstrained		
baseline	1,212 (14.2%)	1,414 (7.1%)
factored	386 (6.6%)	442 (3.1%)

Table 3: OOV ratios in the development test data (half of the WMT 2015 development data).

FinnWordNet also includes a bilingual thesaurus based on the linked Finnish WordNet (Niemi and Lindén, 2012). The HFST tools provide a convenient interface for querying this resource with inflected word forms. We applied this external resources as yet another module for handling OOV words in the input. For this we used the XML

[3] http://code.google.com/p/word2vec/

[4] http://www.ling.helsinki.fi/en/lt/research/finnwordnet/

[5] http://www.ling.helsinki.fi/en/lt/research/finnwordnet/download.shtml#hfst

	OOVs	synonyms
constrained (factored)		
word2vec	626	6,260
- covered by phrase table	371	968
WordNetSyn	318	17,742
- covered by phrase table	262	1,380
unconstrained (factored)		
word2vec	210	2,100
- covered by phrase table	140	480
WordNetSyn	67	2,883
- covered by phrase table	66	361

Table 4: Synonyms extracted from WordNet and word2vec word embeddings for OOVs in the development test data.

markup functionality of Moses to provide translations along with the source language input. The lookup usually leads to several alternative translations including repeated entries (see Table 5 for some statistics). We use relative frequencies and an arbitrary chosen weight factor of 0.1 to determine the probability of the WordNet translation option given to the Moses decoder. The bilingual strategy can also be combined with the synonym approach described above. Here, we prefer translations from the bilingual WordNet and add synonyms if no translation can be found. The results on the development test set are shown in Table 2 as well. Note that we could not use XML markup in connection with factored input. There is, to our knowledge, no obvious way to combine non-factored XML markup with factored input.

WordNetTrans	OOVs	translations
constrained (factored)	336	3,622
unconstrained (factored)	78	532

Table 5: Translations extracted for OOVs in the development test data from the bilingual Finnish-English WordNet.

3.4 Untranslated Words

To evaluate the overall impact of our OOV approach, we inspect untranslated Finnish words in 200 random sentences in the Finnish-English test set output and assign these words into several categories. The corresponding counts are presented in Table 6. Inflected forms account for the vast majority of untranslated output, and of these, inflected proper names constitute more than half. Given that the inflection rules in Finnish are highly productive, a focused effort especially on resolving inflected proper names should be able to account for the majority of the remaining untranslated output. However, since only 52 of the 200 inspected sentences contained untranslated output, no major gains in translation quality can be expected.

category	count
Inflected proper name	35
Inflected non-compound form	13
Inflected compound	9
Other	5
Typo	3
Base form	3
Proper name base form	1

Table 6: Categorization of untranslated Finnish words in the Finnish-English system output.

3.5 Final Results

Our results on the 2015 newstest set are shown in Table 7. Our primary system is the unconstrained factored model with pseudo-tokens and WordNet synonyms. Contrastive runs include the phrase-based baselines and constrained settings in factored and non-factored variants. In the human evaluation, the primary system ranked first shared with five other systems, but this cluster of systems was outperformed by one of the online baselines.

system	$BLEU$	TER
unconstrained		
baseline	18.9	0.737
primary	19.3	0.728
constrained		
baseline	15.5	0.780
factored	17.9	0.749

Table 7: Our final systems tested with the newstest 2015 data set (lowercased BLEU).

4 English-to-Finnish with OPUS

The main purpose of running the other translation direction was to test the impact of additional training data on translation performance. Once again, we simply used the entire database of English-Finnish parallel data sets provided by WMT and OPUS and tested a straightforward phrase-based model without any special treatment and language-specific tools. Again, we relied on lowercased models and used standard procedures to train and tune model parameters. The results are shown in Table 8. In the human evaluation, the primary system ranked first, but was outperformed by both online baselines.

Similar to Finnish-English we can see a strong effect of additional training data. This is not surprising but re-assuring that even noisy data from distant

system	$BLEU_{dev}$	$BLEU$	TER
constrained	12.7	10.7	0.842
unconstrained	15.7	14.8	0.796

Table 8: English-Finnish translation with (*unconstrained*) or without (*constrained*) OPUS (lowercased BLEU and TER on newstest 2015; $BLEU_{dev}$ on development test data).

Feature	Reference	System	Difference
Case Nom	3701/10289	4739/9996	**+11.44pp**
Person Sg3	1620/3947	1991/3867	**+10.44pp**
Mood Ind	2216/3947	2461/3867	**+7.50pp**
Tense Prs	1259/3947	1470/3867	**+6.12pp**
Voice Act	3388/3947	3414/3867	**+2.45pp**
Punct	2874/19772	2283/20004	**+2.38pp**
Infinitive 1	274/3947	352/3867	**+2.16pp**
Unknown	1239/19772	1611/20004	**+1.79pp**
Tense Prt	957/3947	991/3867	**+1.38pp**
Pers pron	344/10289	453/9996	**+1.19pp**
Case Gen	2637/10289	2050/9996	**-5.12pp**
Pcp Prs	227/3947	87/3867	**-3.50pp**
Cmp Pos	1917/10289	1546/9996	**-3.17pp**
Pcp Prf	647/3947	515/3867	**-3.07pp**
Person Pl3	403/3947	277/3867	**-3.05pp**
Voice Pass	436/3947	317/3867	**-2.85pp**
Case Ela	517/10289	219/9996	**-2.83pp**
Uppercase	3126/19772	2624/20004	**-2.69pp**
Prop noun	1675/10289	1399/9996	**-2.28pp**
Case Ine	771/10289	530/9996	**-2.19pp**

Table 9: The ten most over- and under-represented morphological features in the system output as compared to the reference translation. The relative frequency of each feature is calculated with respect to the token count of the word category which exhibits it: nouns, adjectives, pronouns and numerals for case and number, verbs for features like person and tense, and all tokens for generic features like unknown and uppercase.

domains can contribute significantly when training statistical MT models with scarce in-domain training data. The overall quality, however, is still poor as our manual inspections reveal as well. The following section discusses some of the issues that may guide developments in the future.

4.1 Morphological Richness

To study how well the morphological variation is handled in the English-to-Finnish translation direction, we compare the morphological richness of the system output and reference translations. Most over- and under-represented morphological features are shown in Table 9.

For words inflecting in case and number, the nominative case is highly over-represented in the system output. As the nominative case corresponds to the basic form of a word (canonical form), presumably the translation system fails to produce correct inflections when translating from English to Finnish and uses the basic form too often. This naturally leads to the under-representation of other cases. From Table 9 we can see that, e.g., the genitive, elative and inessive cases are underrepresented in the system output. Similar behavior can be seen with verb features as well. Frequent verb inflections are over-represented to the detriment of rarer variants. For example, third person singular and first infinitive (canonical form) are over-represented compared to other persons. Additionally, active forms dominate over passive, and present and past tenses over participial counterparts. Both of these word categories indicate that the morphological variation is weaker in the system output than in reference translations. This shows that the system is not fully able to account for the rich morphology of the Finnish language.

From Table 9 we can also notice several features not directly related to morphology. As expected, the proportion of words not recognized by the Finnish morphological analyzer (*Unknown* row) is higher in system output than in reference translations. This likely reflects words passed through the pipeline untranslated. Moreover, system output has more punctuation tokens and less uppercased words, which is due to the re-capitalization procedure we apply on the originally lowercased output of the decoder.

5 Conclusions

This paper presents baseline systems for the translation between Finnish and English in both directions. Our main effort refers to the inclusion of additional training data and morphological pre-processing for the translation from Finnish to English. We can show that additional noisy and unrelated training data has a significant impact on translation performance and that morphological analyses is essential in this task. Our models perform well relative to other systems submitted to WMT but still underperform in quality as manual inspections reveal. The challenge of translating from and to morphologically rich languages with scarce domain-specific resources is still far from being solved with currents standard technology in statistical machine translation and provides an exciting research field for future work.

References

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pages 4585–4592.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of NAACL*, pages 644–648.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674. Association for Computational Linguistics.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL*, pages 690–696.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL*, pages 967–975.

Jenna Kanerva, Juhani Luotolahti, Veronika Laippala, and Filip Ginter. 2014. Syntactic n-gram collection from a large-scale corpus of Internet Finnish. In *Proceedings Baltic HLT*, pages 184–191.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, pages 177–180.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST tools for morphology — an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, volume 41 of *Communications in Computer and Information Science*, pages 28–47. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Jyrki Niemi and Krister Lindén. 2012. Representing the translation relation in a bilingual WordNet. In *Proceedings of LREC*, pages 2439–2446.

Jyrki Niemi, Krister Lindén, and Mirka Hyvärinen. 2012. Using a bilingual resource to add synonyms to a wordnet: FinnWordNet and Wikipedia as an example. In *In Proceedings of the 6th International Global WordNet Conference (GWC 2012)*, pages 227–231, Matsue, Japan.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pages 160–167.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of LREC*, pages 1850–1855.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Procedings of LREC*, pages 2214–2218.

Jörg Tiedemann. 2014. Improved text extraction from PDF documents for large-scale natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 1 of *Lecture Notes in Computer Science LNCS 8403*, pages 102–112. Springer.