Europarl: A Parallel Corpus for Statistical Machine Translation

Europarl: A Parallel Corpus for Statistical Machine Translation

Philipp Koehn
School of Informatics
University of Edinburgh, Scotland
pkoehn@inf.ed.ac.uk

Abstract

We collected a corpus of parallel text in 11 languages from the proceedings of the European Parliament, which are published on the web[1]. This corpus has found widespread use in the NLP community. Here, we focus on its acquisition and its application as training data for statistical machine translation (SMT). We trained SMT systems for 110 language pairs, which reveal interesting clues into the challenges ahead.

1   Introduction

In many ways, progress in natural language research is driven by the availability of data. This is particularly true for the field of statistical machine translation, which thrives on the emergence of large quantities of parallel text: text paired with its translation into a second language.

The source for these parallel texts are often multinational institutions such as the United Nations or the European Union, but also the governments of multilingual countries such as Canada (French, English) or the Hong Kong (English, Chinese). Harvesting these resources allowed the continued improvement of statistical machine translation systems that challenge the state of the art in MT for many language pairs.

One contribution to this endeavour is the acquisition of the Europarl corpus, which we describe in this paper. It is a collection of the proceedings of the European Parliament, dating back to 1996. Altogether, the corpus comprises of about 30 million words for each of the 11 official languages of the European Union: Danish (da), German (de), Greek (el), English (en), Spanish (es), Finnish (fi), French (fr), Italian (it), Dutch (nl), Portuguese (pt), and Swedish (sv). With the enlargement of the European Union to 25 member countries in May 2004, the European Union has begun to translate texts into even more languages.

---

[1] Available online at http://www.statmt.org/europarl/

We collected the Europarl corpus mainly to aid our research in statistical machine translation, but since we made it available in its initial release in 2001, it has been used for many other natural language problems: word sense disambiguation, anaphora resolution, information extraction, etc.

This paper describes the acquisition of the corpus and its application to the task of statistical machine translation. We used the corpus to build 110 machine translation systems for all the possible language pairs. The resulting systems and their performances demonstrate the different challenges for statistical machine translation for different language pairs.

The field has been dominated by efforts to build MT systems for a handful of languages (Arabic, Chinese, German, French, Spanish) into English. We hope that this contribution stimulates research on non-traditional language pairs.

2   Corpus Collection

Acquisition of a parallel corpus for the use in a statistical machine translation system typically takes five steps:

- obtain the raw data (e.g., by crawling the web)

- extract and map parallel chunks of text (document alignment)

- break the text into sentences (sentence splitting)

- prepare the corpus for SMT systems (normalisation, tokenisation)

- map sentences in one language sentences in the other language (sentence alignment)

In the following, we will describe in detail the acquisition of the Europarl corpus from the website of the European Parliament. These proceedings are published in all of the 11 (former) official languages of the European Union. This means that we can not only extract a conventional parallel corpus, but

```
<CHAPTER ID=1>
Resumption of the session
<SPEAKER ID=1 NAME="President">
I declare resumed the session of the European Parliament ...
<P>
Although, as you will have seen, the dreaded 'millennium bug' ...
```

Figure 1: Format of the released corpus: beginning of file `de-en/en/ep-00-01-17.txt`, the English half of the German–English corpus from January 17, 2000.

a multilingual corpus of 11 languages, or 10 parallel corpora for each language.

2.1 Crawling

The website of the European Parliament[2] provides the Proceedings of the European Parliament in form of HTML files. At the time of our most recent crawl, each file contains the utterances of one speaker in turn. The format has changed this year. The URL for each file contains relevant information for identification, such as its language, the day and number of the thread of discussion and number of the utterance.

Crawling this web resource with a web spider is done by starting at an index page and following certain links based on inclusion and exclusion rules. Since the corpus consists of many small parts, the crawling process is time consuming. Per language, it took several days to obtain the roughly 80,000 files each. Although such crawls are slow, it is typically easier than contacting directly the technical staff of the website and negotiate a transfer procedure.

Usually, there are also copyright concerns, although less so for information from government sources. The European Parliament web site states: "Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged." Such liberal copyright policy can not necessarily expected. Often a longer legal process is required to get permission and access to the data.

Besides identifying sources for parallel corpora manually, it is also possible to mine the web for such data. Resnik [1999] proposes such a system, called STRAND.

2.2 Document Alignment

Each sitting of the European Parliament covers a number of topics. A first step is to identify the texts belonging to each topic, and matching these between languages. To obtain the maximum amount of data, we match these topics for each of the language pairs.

Large data collections such as the Proceedings of the European Parliament are created over the period of many years, often with changing formatting standards and other sources of error. For instance, part of the "English" part of the proceedings contain actually French texts (21–24 May 1996) at the time of our crawl.

The extraction of relevant text from noisy HTML is a cumbersome enterprise that requires constant refinement and adaptation. We process the HTML data with a Perl program that uses pattern matching to detect and extract the identity of the speaker and her statements including paragraph markers.

There is work on automatically learning systems that extract structured information from web sources or other forms of unstructured data. This task is called wrapper induction. See for instance work by Muslea et al. [1999]. For a single data source, however, a manual approach is often more efficient.

For each day, we store the data in one file per language with some meta information, as shown in Figure 1.

We created parallel corpora involving English in this format. We also provide corpora in sentence aligned format, which we will describe below. Scripts are provided to generate the other parallel corpora.

The document alignment is done without tokenisation and sentence splitting. The motivation behind this is that these are error prone processes for which multiple standards could be applied, and we do not want to force any specific standard at this step.

2.3 Sentence Splitting and Tokenisation

Sentence splitting and tokenisation require specialised tools for each language. Unfortunately, we do not have such tools available for all the languages under consideration.

One problem of sentence splitting is the ambiguity of the period "." as either a end of sentence marker, or as a marker for an abbreviation. For English, French and German, we semi-automatically created a list of known abbreviations that are typically followed by a period. One clue is a lowercased

Language	Days	Chapters	Speaker Turns	Sentences	Words
Danish (da)	492	4,120	90,017	1,032,764	27,153,424
German (de)	492	4,119	90,135	1,023,115	27,302,541
Greek (el)	398	3,712	66,928	746,834	27,772,533
English (en)	488	4,055	88,908	1,011,476	28,521,967
Spanish (es)	492	4,125	90,305	1,029,155	30,007,569
French (fr)	492	4,125	90,335	1,023,523	32,550,260
Finnish (fi)	442	3,627	81,370	941,890	18,841,346
Italian (it)	492	4,117	90,030	979,543	28,786,724
Dutch (nl)	492	4,122	90,112	1,042,482	28,763,729
Portuguese (pt)	492	4,125	90,329	1,014,128	29,213,348
Swedish (sv)	492	3,627	81,246	947,493	23,535,265

Table 1: Size of the released corpus (version 2). The numbers of sentences and words is after tokenisation and sentence-alignment with English (or German, in the case of English).

word following a period ("ca. three thousand men"), which indicates an abbreviation and not an end of a sentence.

There has been extensive work on empirical methods to learn sentence breaking. See for instance the work on SATZ [Palmer and Hearst, 1997]. Various machine learning methods can be applied to this problem, such as decision trees [Riley, 1989] and maximum entropy [Reynar and Ratnaparkhi, 1997].

Issues with tokenisation include the English merging of words such as in "can't" (which we transform to "can not"), or the separation of possessive markers ("the man's" becomes "the man 's"). We do not perform any specialised treatment for other languages than English at this point. In future, we would like to employ a tokenisation scheme that matches the Penn treebank standard. Currently, our provided scripts allow external tokenisation methods.

For training a statistical machine translation system, usually all words are lowercased to eliminate the differences between different spelling of words depending on their occurrence at the beginning of a sentence (*The*), in the middle (*the*), or in a headline (*THE*). A more sophisticated approach is truecasing, which allows the distinction names (*Mr. Black*) and regular words (*black*).

2.4 Sentence Alignment

Sentence alignment is usually a hard problem, but in our case it is simplified by the fact that the texts are already available in paragraph aligned format. Each paragraph consists typically of only 2–5 sentences.

If the number of paragraphs of a speaker utterance differs in the two languages, we discard this data for quality reasons. The alignment of sentences in the corpus is done with an implementation of the algorithm by Gale and Church [1993]. This algorithms tries to match sentences of similar length in sequence and merges sentences if necessary (e.g. two short sentences in one language to one long sentence in the other language), based on the number of words in the sentence. Since there are so few sentences per paragraph, alignment quality is very high.

There is considerable work on better sentence alignment algorithms. One obvious extension is to not only consider sentence length, but also potential word correspondences within sentence pairs. Work by Melamed [1999] is an example for such an approach.

The sentence aligned data is stored in one file per day per language, so that lines with the same line number in a file pair are mappings of each other. The markup from the document aligned files is stripped out. In the current release, the size of the sentence aligned corpus is roughly 30 million words in one million sentences per language. Fore more detailed statistics, see Table 1.

2.5 Extraction of a Common Test Set

To allow the comparison of machine translation system, it is necessary not only to define a common training set (as the Europarl corpus), but also a common test set. We suggest to reserve the last quarter of 2000 (November–December 2000) as test set, and to use the rest of the corpus as training data. This portion of the corpus comprises of over one million words in over 40,000 sentences.

To be able to compare system performance on different language pairs, we also extracted a set of sentences that are aligned to each other across all 11 languages. Figure 2 is one of the sentences from this collection.

Danish: det er næsten en personlig rekord for mig dette efterår .
German: das ist für mich fast persönlicher rekord in diesem herbst .
Greek: πρόκειται για το προσωπικό μου ρεκόρ αυτό το φθινόπωρο .
English that is almost a personal record for me this autumn !
Spanish: es la mejor marca que he alcanzado este otoño .
Finnish: se on melkein minun ennätykseni tänä syksynä !
French: c ' est pratiquement un record personnel pour moi , cet automne !
Italian: e ' quasi il mio record personale dell ' autunno .
Dutch: dit is haast een persoonlijk record deze herfst .
Portuguese: é quase o meu recorde pessoal deste semestre !
Swedish: det är nästan personligt rekord för mig denna höst !

Figure 2: One sentence aligned across 11 languages

Note that this data is also lowercased, which is not done for the released sentence aligned data. Alternatively, true casing could be applied, although this is a more difficult task.

2.6 Releases of the Corpus

The initial release of this corpus consisted of data up to 2001. The second release added data up to 2003, increasing the size from just over 20 million words to up to 30 million words per language. A forthcoming third release will include data up to early 2005 and will have better tokenisation. For more details, please check the website.

3 110 SMT Systems

The prevailing methodology in statistical machine translation (SMT) has progressed from the initial word-based IBM Models [Brown et al., 1993] to current phrase-based models [Koehn et al., 2003]. To describe the latter quickly: When translating a sentence, source language phrases (any sequences of words) are mapped into phrases in the target language, as specified by a probabilistic phrase translation table. Phrases may be reordered, and a language model in the target language supports fluent output.

The core of this model is the probabilistic phrase translation table that is learned from a parallel corpora. There are various methods to train this, several start with a automatically obtained word alignment and then collect phrase pairs of any length that are consistent with the word alignment.

Decoding is a beam search over all possible segmentation of the input into phrases, any translation for each phrase, and any reordering. Additional component models aid in scoring alternative translations. Translation speed in our case is a few seconds per sentence.

Fuelled by annual competitions and an active research community, we can observe rapid progress in the field. Due to the involvement of US funding agencies, most research groups focus on the translation from Arabic to English and Chinese to English. Next to text-to-text translation, there is increasing interest in speech-to-text translation.

Most systems are largely language-independent, and building a SMT system for a new language pair is mostly a matter of availability of parallel texts. Our efforts to explore open-domain German–English SMT led us to collecting data from the European Parliament. Incidentally, the existence of translations in 11 languages now enabled us to build translation systems for all 110 language pairs.

Our SMT system [Koehn et al., 2003] includes the decoder Pharaoh [Koehn, 2004], which is freely available for research purposes[3]. Training 110 systems took about 3 weeks on a 16-node Linux cluster. We evaluated the quality of the system with the widely used BLEU metric [Papineni et al., 2002], which measures overlap with a reference translation.

We tested on a 2000 sentences held-out test set, which is drawn from text from sessions that took part the last quarter of the year 2000. These sentences are aligned across all 11 languages, so when translation the, say, French sentences into Danish, we can compare the output against the Danish set of sentences. The same test set was used in a shared task at the 2005 ACL Workshop on Parallel Texts [Koehn, 2005].

The scores for the 110 systems are displayed in Table 2. According to these numbers, the easiest translation direction is Spanish to French (BLEU score of 40.2), the hardest Dutch to Finnish (10.3).

---

[3]Available online at `http://www.isi.edu/licensed-sw/pharaoh/`

Source	Target Language										
Language	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus
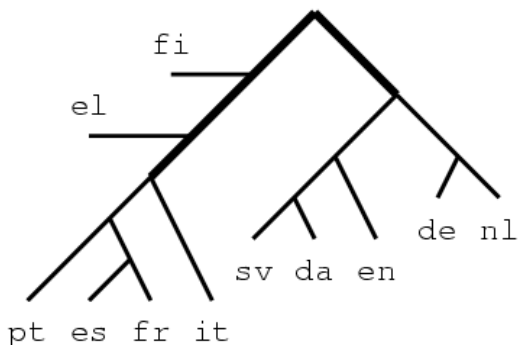
Figure 3: Clustering of languages based on system scores: Language families emerge

Language	From	Into	Diff
Danish (da)	23.4	23.3	0.0
German (de)	**22.2**	**17.7**	**-4.5**
Greek (el)	23.8	22.9	-0.9
English (en)	**23.8**	**27.4**	**+3.6**
Spanish (es)	26.7	29.6	+2.9
French (fr)	26.1	31.1	+5.1
Finnish (fi)	19.1	12.4	-6.7
Italian (it)	24.3	25.4	+1.1
Dutch (nl)	19.7	20.7	+1.1
Portuguese (pt)	26.1	27.0	+0.9
Swedish (sv)	24.8	22.1	-2.6

Table 3: Average translation scores for systems when translating *from* and *into* a language. Note that German (de) and English (en) are similarly difficult to translate *from*, but English is much easier to translate *into*.

4 Language Clustering

Intuitively, languages that are related are easier to translate into each other. We can underscore this with our SMT system scores. When clustering languages together based on their translation score, the 11 languages group together roughly along the lines of their language families, as shown in Figure 3.

One the one side, you can find the Romance languages Spanish, French, Portuguese and Italian, on the other side the Germanic languages Danish, Swedish, English, Dutch and German. The close languages Danish and Swedish, as well as Dutch and German are group together first. The graph is not perfect: One would suspect Spanish and Portuguese to be joined first, but Spanish is first joined with French.

The clustering algorithm greedily groups languages together that translate into each other most easily. In the first step, Spanish and French are grouped together, since they have the highest translation score (38.4 and 40.2). In the next step Portuguese is added (37.9 and 35.9 with Spanish, 39.0

and 35.3 with French). Always, the two clusters of languages are joined that have the highest average translation score. A bias term of $-|c_1| \times |c_2|/2$ is added to the score to bias toward the emergence of smaller clusters ($|c|$ is the size of the cluster c).

5 Translation Direction

Some language are more difficult to translate *into* than *from*. See Table 3 for details on this. The average score for systems that translate from German into the each of the other 10 languages is 22.2, very similar for systems translating from English, 23.8. However, the scores for translating *into* these language is vastly different: 17.7 for German vs. 27.4 for English.

One apparent reason for the difficulty of translating into a language is morphological richness. Noun phrases in German are marked with case, which
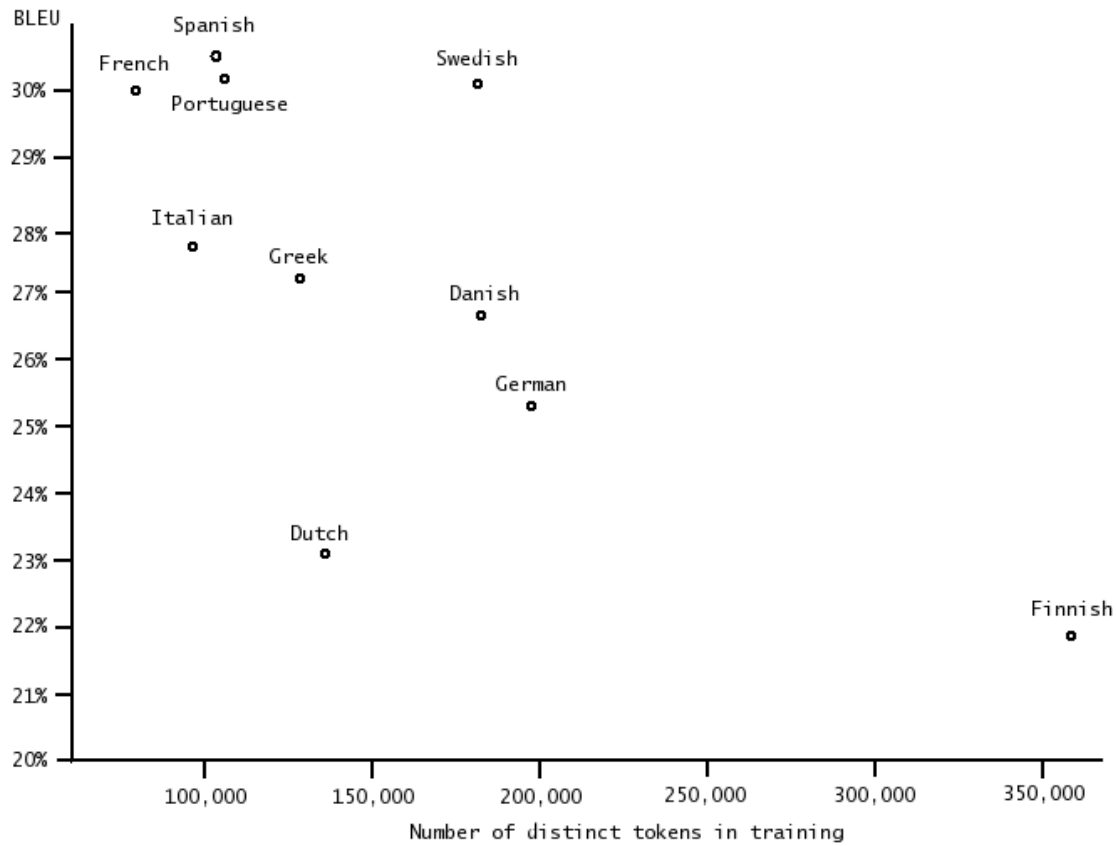
Figure 4: Vocabulary size vs. BLEU score when translating into English (which has about 65,000 distinct word forms)

manifests itself as different word endings at determiners, adjectives and nouns. Generating the right case markings is hard, especially since nothing in the translation model keeps track of the role of noun phrases and the trigram language model is fairly weak in this respect, since it only considers a three word window.

The poor performance of systems involving Finnish can also partly be attributed to its morphology, which is very agglutinative: Some elements that form individual words in English (determiners, prepositions) are included in the morphology. This increases the size of the vocabulary (the Finnish vocabulary is about five times as big as the English), leading to sparse data problems when collecting statistics for word and phrase translation. See Figure 4 for a comparison of BLEU scores when translating into into English and vocabulary size.

Intuitively, translating from an information-rich into an information-poor language is easier than the other way around. Researchers have made similar observations about the better performance of Arabic–English SMT systems vs. Chinese–English

SMT systems, that are trained on similar amount of training data and tested on news wire: Translating from Arabic with its rich morphology is easier than translating from Chinese, which is even more frugal than English, often lacking determiners and plural or tense markers.

Note that translating into English is among the easiest. However, since the research community is primarily occupied with translation into English, interesting problems associated with translating into morphologically rich languages have largely been neglected.

6 Back Translation

The quality of machine translation systems is difficult to assess. This is especially true for monolingual speakers, who only know one language. When mainstream journalists report on the progress of machine translation systems, they frequently resort to a seemingly clever trick: They use a MT system to translate a sentence from English into a foreign language, and then use a reverse MT system to translate the sentence back into English. They then judge the

Language	From	Into	Back
da	28.5	25.2	56.6
de	25.3	17.6	48.8
el	27.2	23.2	**56.5**
es	30.5	30.1	52.6
fi	21.8	13.0	44.4
it	27.8	25.3	49.9
nl	23.0	21.0	46.0
pt	30.1	27.1	**53.6**
sv	30.2	24.8	54.4

Table 4: Scores for mono-directional systems and back translation: Translating from English to Greek (system score 27.2) and back to English (system score 23.2) results in a BLEU score of 56.5 for the combined translation. The score is higher than for the combination English–Portuguese–English (53.6), although the mono-directional systems are better (30.1, 27.1).

quality of the MT systems by how well the English sentence is preserved.

This method is inspired by an urban legend involving a pair of MT systems between Russian and English. The legend proclaims that once someone fed a English–Russian MT system the bible verse *"The spirit is willing, but the flesh is weak."* When back translating the sentence with the Russian–English system, the system returned *"The vodka is good but the meat is rotten."*

How well does back translation indicate the translation performance of the MT systems involved? As Table 4 shows, not much.

First of all, while one would suspect a degradation of the quality of a sentence when translated into a foreign language, and a further degradation when translated back, the BLEU scores tell a different story: For instance, the quality of the English–Greek system is 27.2 and 23.2 for the Greek–English system. However, translating the test set from English into Greek and back into English, gives a BLEU score of 56.5, much higher than either system.

Note that this high score is an artifact of how the BLEU score works: It measures overlap with a reference translation. In the mono-directional systems the reference translation is a human translation. While the system output may be correct, the system may get punished for valid translation choices that differ from the ones by the human. In back translation, however, we compare against exactly the input sentence, which will be easier to match.

The more interesting point of Table 4 is: The back translation scores do not correlate well with the mono-directional system scores. Again, the English–Greek–English combination has system scores of 27.2 and 23.2, and a back translation score of 56.5. This is higher than 53.6, the score for the English–Portuguese–English combination, which has better mono-directional system scores: 30.1 and 27.2.

In conclusion, back translation does not only provide a false sense of the capabilities of MT systems, it is also a lazy and flawed method to compare systems. Back translation unfairly benefits from the ability to reverse errors, which only show up in the foreign language. To drive the point home: a system pair that does nothing, meaning, leaving all English words in place will do perfectly in back translation, while being utterly useless in practise.

7 Conclusions

We described the acquisition of the Europarl corpus and its application in building statistical machine translation systems for 110 language pairs, maybe the largest number of machine translation systems built within three weeks, and the first serious effort at building such a system for, say, Greek to Finnish. Some sample output is in Figure 5.

The widely ranging quality of the different SMT systems for the different language pairs demonstrate the many different challenges for SMT research, which we have only touched upon. The field's primary occupation with translating a few languages into English ignores many of these challenges.

Finally, we hope that the availability of resources (corpora, tools) continues to make statistical machine translation an exciting and productive field.

References

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

Gale, W. and Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1).

Koehn, P. (2004). Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas, AMTA*, Lecture Notes in Computer Science. Springer.

Koehn, P. (2005). Shared task: Statistical machine translation for european languages. In *ACL Workshop on Parallel Texts*.

Spanish-English
we all know very well that the current treaties are insufficient and that , in the future , it will be necessary to develop a better structure and different for the european union , a structure more constitutional also make it clear what the competences of the member states and which belong to the union .
messages of concern in the first place just before the economic and social problems for the present situation , and in spite of sustained growth , as a result of years of effort on the part of our citizens .
the current situation , unsustainable above all for many self-employed drivers and in the area of agriculture , we must improve without doubt .
in itself , it is good to reach an agreement on procedures , but we have to ensure that this system is not likely to be used as a weapon policy .
now they are also clear rights to be respected .
i agree with the signal warning against the return , which some are tempted to the intergovernmental methods .
there are many of us that we want a federation of nation states .

Finnish-English
the rapporteurs have drawn attention to the quality of the debate and also the need to go further : of course , i can only agree with them .
we know very well that the current treaties are not enough and that in future , it is necessary to develop a better structure for the union and , therefore perustuslaillisempi structure , which also expressed more clearly what the member states and the union is concerned .
first of all , kohtaamiemme economic and social difficulties , there is concern , even if growth is sustainable and the result of the efforts of all , on the part of our citizens .
the current situation , which is unacceptable , in particular , for many carriers and responsible for agriculture , is in any case , to be improved .
agreement on procedures in itself is a good thing , but there is a need to ensure that the system cannot be used as a political lyömäaseena .
they also have a clear picture of the rights of now , in which they have to work .
i agree with him when he warned of the consenting to return to intergovernmental methods .
many of us want of a federal state of the national member states .

Figure 5: Sample output of the trained systems

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL).*

Melamed, D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Muslea, I., Minton, S., and Knoblock, C. (1999). A hierarchical approach to wrapper induction. In Etzioni, O., Müller, J. P., and Bradshaw, J. M., editors, *Proceedings of the Third International Conference on Autonomous Agents (Agents'99)*, pages 190–197, Seattle, WA, USA. ACM Press.

Palmer, D. D. and Hearst, M. A. (1997). Adaptive multilingual sentence boundary detection. *Computational Linguistics*, 23(2):241–267.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL).*

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the International Conference of the Association of Computational Linguistics.*

Reynar, J. C. and Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Firth Conference on Applied Natural Language Processing*, pages 803–806.

Riley, M. D. (1989). Some applications of tree-based modeling to speech and language. In *Proceedings of the DARPA Speech and Language Technology Workshop*, pages 339–352.