The Sentence-Aligned European Patent Corpus

Wolfgang Täger
European Patent Office
wtaeger@epo.org

Abstract

This paper describes the creation and the content of the Sentence-Aligned European Patent Corpus. The corpus contains more than 130 million sentence pairs for 6 European languages.

With more than 76 million sentence pairs, to our knowledge, the EN-DE sub corpus is the largest bilingual sentence-aligned corpus. For other language pairs, work has started to obtain sub corpora of similar size. The error rate of sentence alignment was very low even in the absence of language specific resources.

1 Translations in the patenting process

Patent protection is granted by Patent Offices with temporal and territorial limitations. Each Patent Office has a specific language regime allowing patents to be filed in one or several languages. To obtain broad geographic coverage, it is therefore necessary to translate patent applications and to file these translated patents with other Patent Offices. This may be the case for claiming priority, for entering a national or regional phase in the PCT procedure, or in Europe for validating a European Patent in Member States.

Details of the language regime of the European Patent Office (EPO) may be found in the European Patent Convention (EPC), in particular in Art. 14, 65, 67, 149a and 153.
Similar legal provisions exist for other Patent Offices and for the Patent Cooperation Treaty PCT (http://www.wipo.int/pct). In recent research on sentence alignment of patent translations, the focus was either on priority documents (Utiyama and Isahara, 2007) or on PCT documents (Lu et al., 2010). In both cases difficulties were reported due to the fact that the patent pairs frequently were not true translation pairs, and heavy filtering was applied to reduce error rates, reducing the number of alignments to approximately 30% of the original number.

With European Patents (EP) the EPO is in a more comfortable situation of having real translations in the case of Art. 14 (trilingual claims) and Art. 65 (post-grant national translations).

As an example you may find EP1234555B1 on espacenet™ (http://worldwide.espacenet.com/numberSearch) with an EN-DE-FR set of claims and an EN description, and then translations ES2279853T3 and DE60217981T2 under “also published as”. The kind codes T3 for Spain and T2 for Germany indicate that these are translations of the EP-B1.

In this example, we therefore have a trilingual set EN-DE-ES of descriptions and claims in 4 languages including FR. In the following we will refer to this as the **POST-GRANT corpus**.

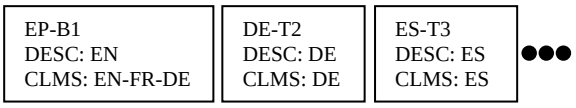


Figure 1. POST-GRANT translations

Patent applicants may under certain conditions enjoy priority rights from one or several earlier applications disclosing *the same invention*. In such cases novelty and inventiveness are examined based on this earlier date. It is however important that the European application (EP-A) is not necessarily a true translation of the priority document(s).

As priority documents may be drafted in any language, the EPO *may* request translation of the priority document, in particular when the priority date is relevant for the decision to grant. The priority document and its translation can be used as a true pair of translated patents. We refer to

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 177–184
Leuven, Belgium, May 2011

this as the **PRIODOC-PRIOTRAN corpus**. In most cases however no PRIOTRAN exists.

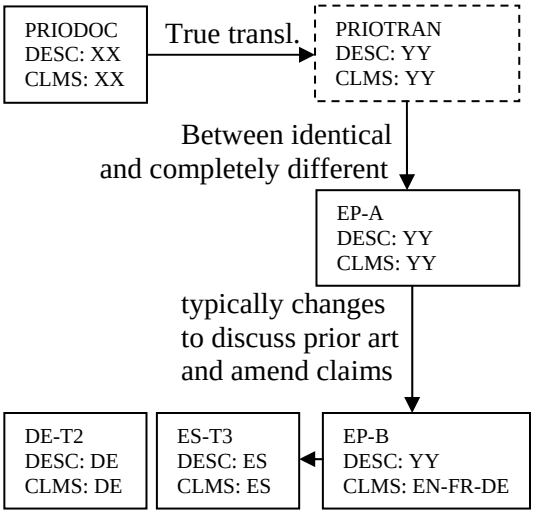


Figure 2. Workflow Priority, A-, B-,T-docs.

As an example you may find in the European patent register (https://register.epo.org/espacenet/regviewer) for EP1000000 a priority document in Dutch (XX=NL) and a translation into EN (YY=EN). In this specific case the European application EP1000000A happens to be identical to the PRIOTRAN, but there is no guarantee that this is always the case.

PRIODOC against EP-A (**PRIODOC-A corpus**) or against the granted patent or its translations (**PRIODOC-B/T corpus)** is therefore considered either as a noisy corpus or as a comparable corpus.

According to Art. 67, applicants may request provisional protection based on the EP-A document. For this the applicant might have to file translations of the claims (**PROVISIONAL CLAIMS corpus**). We did not follow this route so far. Similarly further translations such as translations of titles (**TITLE corpus**) and abstracts (**ABSTRACT corpus)** were not considered yet because their size is smaller.

For translations via the PCT route, one may distinguish between:

- EPO as International Search Authority (ISA), which is essentially the same case as discussed above,
- EPO as International Preliminary Examination Authority; however in most cases the EPO then was also the ISA, so no further translation is required,
- and EPO as Regional Office.

In the latter case (**PCT corpus**), most translations stem from non European languages. From WIPO's Patentscope one may derive that the number of International Applications sent to the Spanish Office is only 11 000 patents of which the EPO / USPTO received 4 700 / 2 700 as Regional Office, respectively. Compared to 286 000 Spanish T3 documents, this amount is low. Even the 92 000 International Applications in German with national phase in the United States are small compared to 482 000 DE-T2, and some overlap is likely.

As translating between European languages enjoys higher priority at the EPO, we have not yet studied in detail the PCT corpus. However it is likely that this corpus suffers from the amendments during the procedure. Even though true translations may exist (like PRIOTRAN), in practice it might be difficult to retrieve those true translations. MT researchers are more likely to obtain PRIODOC-A pairs.

In total we have the following corpora types:

Corpus	True translat.	Comment
POST-GRANT	Yes	Multilingual EN, DE, FR, ES, …
PRIODOC-PRIOTRAN	Yes	Smaller bilingual corpora for each Office.
PRIODOC-A	Not sure	Large noisy bilingual corpora for each Office.
PRIODOC-B/T	No	Large multilingual corpus, but two noise sources.
PROVISIONAL CLAIMS	Yes	Multilingual with fewer languages, only claims.
TITLE	Yes	Multilingual, but a title is not even a single complete sentence per language.
ABSTRACT	Yes/No	E.g. at WIPO.
PCT	Yes/No	See e.g. http://pct.intl-patent.com/en/

Table 1. Corpora types.

2 Corpus collection

In order to validate a European Patent in a specific Member State, the applicant typically has to file translations to National Patent Offices. In 2008 however, the London Agreement (http://www.epo.org/law-practice/legal-texts/london-agreement.html) came into force, in which several countries (currently 16 out of 38) waived their right to demand translations of descriptions and/or claims. Other countries such as Italy, Spain and Austria still request full translations.

The number of post-grant translations of European Patents depends on a variety of factors for the patent applicant such as when the country joined the European Patent Convention (http://www.epo.org/about-us/organisation/member-states/date.html), economic importance of a specific country for the specific invention, costs for translation, validation and renewal fees, estimated value of the patent and others.

Year	Member States	Total Patents	Events
1977	7	0	EPO set up
Jan. 1990	14	125 796	
Jan. 2000	19	479 885	
Apr. 2008	34	893 252	London Agreement
Feb. 2011	38	1 054 489	
?			Unitary Patent

Table 2. History of EPC accession and patents.

Many European Patents (estimated >250 000) were validated in Germany, France, UK, Italy, Spain, the Netherlands and Belgium. Large quantities (at least 50 000, some around 250 000) were validated in Sweden, Denmark, Switzerland, Austria, Greece, Turkey, Portugal and Poland. Fewer patents were validated in the remaining countries.

The translations were filed at the respective National Offices. Some of them were kept in paper form only, some were scanned and made publicly available, others were even OCRed and made available in text format.

Together with National Patent Offices of its Member States, the European Patent Office is now digitalizing additional translations to increase the corpus size.

3 Document preprocessing

For translations filed and OCRed at National Offices, there is no guarantee that the same format is used. XML and Unicode are widespread now based on WIPO standards, but the EPO also received texts in Latex or without any markup.

If the OCRing process was not adapted to the layout, line numbering and headers and footers may appear somewhere in the text. Even worse, the OCR may eliminate layout information identifying the beginning and ending of the different parts (title, abstract, description, claims).

Paragraphs identified by the OCR process may significantly differ from paragraphs of the European Patent, e.g. a list may be represented as one or as many paragraphs.

Chemical formulas may be written in XML or embedded in image format. Old translations have been written by typewriters. Some pages contain handwritten remarks, corrections or stamps.

High quality OCR is therefore required which is more costly. Some National Offices already have digitalized their translations in high quality text format, in particular the German and the Spanish Patent Office. These texts were converted into XML-Unicode. For other Offices, re-OCRing might be necessary.

The costs of high quality OCR are below 2€ per patent which may be compared with the translation costs of around 1 200€ per patent.

4 Sentence boundary detection

For sentence boundary detection different methods are known such as rule-based regular expressions, data-driven machine learning algorithms and syntax-based methods based on part-of-speech taggers.

For the sake of simplicity, we used regular expressions. The most important step was to identify the abbreviations used in the patents. For this the texts were scanned for sequences of a dot followed by anything which may not be a start of a sentence, e.g. a comma or a lowercase letter.

Another method is to identify sequences of a dot followed by a word and another dot, which frequently reveal abbreviations used in citations of prior art such as “Acad. Sci.”.

The abbreviation candidates obtained by such processes are then manually verified and integrated into a regular expression.

In addition, hints showing that a dot is a sentence boundary may be used, e.g. if in an English text a dot is followed by a space and “The”. This rule has priority over the abbreviation list.

5 Sentence alignment

Sentence alignment of patents has been studied previously in less favorable circumstances, e.g. on priority documents, requiring complicated processing with linguistic resources such as dictionaries, stemmers or SMT systems.

In our case, we obtained excellent results without linguistic resources apart from the list of common abbreviations which was obtained semi-automatically as described before.

The algorithm is based on a two-stage Viterbi processing. As a first step, paragraphs are detected and aligned. In the second stage sentences in the paragraphs are detected and aligned.

To improve quality, different methods to identify paragraphs were tested for each patent pair. The one which produced the most equal number of paragraphs in the two patents was selected. The search area is determined around a diagonal which is based on character counts, e.g. if 40% of the source language text is before a paragraph boundary we search for the first paragraph boundary for which more than 40% of the target language text is before the boundary. The search area obtained this way performed much better than a version based on paragraph numbers only.

The scoring function S if an alignment (ds, dt) is plausible depends on the text length score S_{len}, a score S_{invar} derived from translation invariants such as numbers and capital letter sequences, and a penalty p to penalize many-to-one and many-to-many alignments.

$$S_1 = (1-w)S_{len} + w\,S_{invar}$$

$$S = p\,S_1 = p_0^{\max(0,\mathrm{ds}+\mathrm{dt}-2)} S_1$$

Assuming text lengths l_1 and l_2 and a language pair specific constant c, which is the expected text length ratio, e.g. $c = 1.1$ for EN-DE, the scoring function is

$$S_{len} = f(l_1, l_2, c)$$

Gale and Church (1993) assumed a normal distribution for

$$\delta = (l_2 - l_1 c)/\sqrt{l_1 s^2}$$

We did not follow their approach for a variety of reasons. Firstly, the Gauss error function decreases fast, so that outliers which may arise from different encoding, e.g. embedded image vs. text, are given very low scores. Secondly, the formula lacks symmetry

$$S_{len}(l_1, l_2, c) \neq S_{len}\left(l_2, l_1, \frac{1}{c}\right)$$

and does not explain deletions such as (0:1) alignments for which $l_1 = 0$. Therefore we applied an empirical formula:

$$S_{len} = \left(1 - \frac{|l_2 - l_1 c|}{l_2 + l_1 c + 10(c+1)}\right)^{1+\frac{l_2+l_1 c}{200}}$$

which is nearly symmetrical when c is close to one. S_{invar} is calculated from co-occurrence and permutation of items such as numbers which are expected to be translation invariants.

The weighting factor w is calculated by:

$$w = \frac{300n}{300n + l_1 + l_2}$$

If there are many numbers n in a text fragment pair in relation to the text lengths, then w becomes close to 1, giving a high weight to S_{invar}. Due to the high frequency of numbers in patent texts, the alignment algorithm relies to a large extent on the co-occurrence and order of numbers in patents. Note that for descriptions the average length of a sentence is 183 characters, so for a typical sentence if there is one number in each language ($n = 2$) the weighting factor would already put more than 60% weight on S_{invar}. The table below shows the frequency of numbers in sentences from patent descriptions:

Numbers	Percentage
0	37%
1	16%
2	13%
3	9%
4	7%
5	5%
6	3%
7	2%
8	2%
9	1%
≥10	4%

The penalty p was used to avoid the algorithm preferring longer alignments over shorter ones, for example a (2:2) alignment over two (1:1) alignments. Empirically p_0 was set to 0.8.

The algorithm performed beyond our expectations. For a test sample of 200 pairs, we achieved 99% correct alignments for our largest English-German sub corpus (EN-DE descriptions), the remaining 2 pairs were partial translations. The explanation for this high precision is that there

are so many numbers in patents that it is very unlikely that the algorithm would follow an incorrect path over many sentences.

Ambiguous cases where two Viterbi paths nearly obtain the same result could either be discarded or treated with more complex algorithms, e.g. cognate recognition. The second best score was therefore stored as well, but given the high quality already achieved we actually neither discarded pairs based on second best information, nor did we use cognate recognition.

Contrary to all sentence alignment publications known to the author, we were the first ones to allow many-to-one alignments where many means larger than 50. The following table shows alignments of a sub corpus of EP-B before 2005 aligned with DE-T2 (in 2005 the format changed from SGML to XML).

Alignment ds:dt	Count	Percentage
1:1	51 313 726	90.676%
2:1	1 993 957	3.524%
1:2	1 402 537	2.478%
2:2	648 148	1.145%
1:0	296 319	0.524%
…		
10:1	11 366	0.020%
20:1	1 661	0.003%
50:1	160	0.000%

Table 3. Sent. Alignments EP-B1 vs. DE-T2

The reason why we allowed such alignments is that original and translation did not stem from the same processing, but they were independently scanned, OCRed and encoded. This not only concerns the paragraphs, but also enumeration elements, embedded images, tables and so on.

6 Filtering

Approximately 3% of the sentence pairs scored lower than 0.5 ($S < 0.5$) and were discarded. As one-to-many alignments are penalized by the scoring function, the extreme cases 1:n with n>3 were therefore eliminated from the result, but they contributed to find the correct alignment path.

As claims are numbered, claim alignment could easily be obtained by pairing claims with the same number, but we preferred to apply the same algorithm to detect and eliminate errors, and to break down alignment even further, e.g. to enumerations and characterizing part of a claim.

The only case with a high error rate was due to bad OCRing of Portuguese documents. All other alignments achieved 98..99% precision.

7 Parallel corpora: languages, size, quality

7.1 Overview

Table 4 shows the sub corpora obtained at the EPO until now, indicating separately descriptions D and claims C.

Sub corpus	Docs/Lang. used	Sentence pairs[1]
POST-GRANT	EP-B1 EN + DE-T2	59 825 821 D
POST-GRANT	EP-B1 EN + ES-T3	5 768 273 D 137 554 C
POST-GRANT	EP-B1 EN + PT-E	3 782 037 D
POST-GRANT	EP-B1 FR + DE-T2	4 408 895 D
POST-GRANT	EP-B1 FR + ES-T3	1 130 546 D 142 237 C
POST-GRANT	EP-B1 DE + ES-T3	2 285 537 D
POST-GRANT	DE-T2 + ES-T3	6 205 717 D 582 704 C
POST-GRANT	EP-B1 EN + EP-B1 FR	17 089 184 C
POST-GRANT	EP-B1 EN + EP-B1 DE	16 878 483 C
POST-GRANT	EP-B1 FR + EP-B1 DE	14 262 565 C
PRIODOC-PRIOTRAN[2]	IT PRIODOC + EN PRIOTRAN	3 980 298 D 496 417 C
Total		**87 387 124 D** **49 589 144 C**

Table 4: Size of aligned corpora

This corpus only uses a subset of all existing EP, DE, ES, PT and IT documents.

7.2 DE-EN description sub corpus

Our largest sub corpus is based on EN-DE descriptions where the EN description stems from a European Patent in EN, and the DE description from a translation filed at the German Patent and Trademark Office.

This corpus alone comprises nearly 60 million sentence pairs at a very high quality above 99%. It is believed to be the largest available bilingual corpus, exceeding the Chinese-English Patent

[1] Small percentage containing no text only markup
[2] Small parts PRIODOC + EP-A and EP-B1 + IT-T

Corpus of Lu et al. (2010) by a factor of 5 whilst presenting a much higher level of precision.

	Utiyama, JP-EN DESC (2007)	Lu, ZH-EN DESC (2010)	EPO, EN-DE DESC (2011)
Correct	90.0%	89%	99.0%
Partially correct	9.8%	8%	1.0%
Wrong	0.3%	3%	0
Recall	28.5%	31.30%	97%
Size	2 million	12 million	60 million

Table 5. Comparison with previous results.

For better comparison with previous results, the nearly 17 million aligned sentence pairs from EN-DE claims were not included in this table. Hence the EN-DE bilingual sub corpus contains more than 76 million sentence pairs in total.

8 Size impact on SMT

As stated by F. J. Och (2005), doubling a parallel corpus (i.e. the input of the Translation Model TM) is expected to increase absolute BLEU score by 2.5%, whereas doubling the input of the language model LM only yields 0.5%. He presented a BLEU score increase of 4.5% by increasing the LM input from 75 million to 219 billion words, i.e. a factor of nearly 3 000.

At present a parallel corpus of 1 million sentence pairs per language pair was considered a large corpus (e.g. JRC Acquis). If we now have $2^6 = 64$ times more parallel sentence pairs, according to Och's formula a BLEU score increase of 6*2.5%=15% is to be expected, so a much smaller size increase (factor 64 for TM instead of 3 000 for LM) leads to a much higher BLEU score increase than the reported 4.5%.

If Google tried to obtain the missing 10.5% BLEU score increase by further enlarging the language model corpus, according to their own formula they would need to increase the already huge web corpus by a factor $2^{10.5\%/0.5\%} = 2^{21}$. So one would need 2 million times the size of the 219 billion words LM to obtain the same BLEU score increase.

These are dramatic figures, of which we are not sure if they are realistic in our case. Lu et al. reported only 1.08% BLEU score increase for doubling the translation model instead of 2.5% (calculated from 4.66% BLEU score increase when increasing TM by a factor of 20).

Koehn (2002) reported only 1.33% BLEU score increase for doubling the translation model.

Both Lu and Koehn reported decreasing BLEU score gains when more data is used.

Nevertheless it is assumed that a large increase of the TM model significantly outperforms a large increase of the LM for BLEU score.

9 MT considerations

Large patent corpora should significantly improve statistical machine translation quality. However making maximum use of the data will be a challenge on its own.

Besides the computational complexity question reported from Tinsley, Way and Sheridan (2010), the following aspects might be studied:

9.1 Impact of sentence alignment accuracy on translation quality

We obtained very high alignment accuracy because the corpus contains true translations with many numbers. In other cases this may not be available. How would alignment errors affect SMT quality? Should patent families be used, which are not true translations? Which precision/recall trade-off is optimal for SMT quality?

9.2 Translation invariants

Translation invariants are text parts which are not translated. Typical examples are:

- Numbers
- Special characters like Greek letters
- Parentheses and brackets
- Proper Names
- Markup (e.g. XML markup for bold)

For translating the rest of a sentence, the precise content of the translation invariant typically does not have any impact, e.g. it is irrelevant for the rest if a reference number is (5) or (6), if it is Peter or Paul (maybe Mary would be slightly different) or if a part of the text is in bold. In many cases they could be handled as a semantic-free tokens to be copied to the right place in the translated text.

However, current SMT systems may treat them as semantic-bearing tokens, so the n-gram phrase table may have separate items e.g. for "the echo canceller (5) operates" and "the echo canceller (7) operates".

This issue was partly addressed by Huet et al. for WMT 2010. Alas, since we indeed observed permutations between source and target, in particular for languages not following a subject-

verb-object order like DE, the solution proposed there might not work well in our case.

Therefore one might look into the question how to best preserve translation invariants without negative influence on the phrase table.

9.3 Domain aware SMT

There is a widely used international classification system for patents called International Patent Classification (IPC). The classification is hierarchical and extremely detailed.

SMT systems are known to perform better on in-domain texts than on out-of-domain texts. It was reported (Utiyama and Isahara 2007) that training a system using patents from all domains gave better results than just using patents from the same domain, confirming the mantra "more data is better data". Still the question remains how to best combine domains (Dobrinkat and Väyrynen, 2010, and Yasuda et al. 2008) given the specific IPC structure.

10 Conclusion and Outlook

Patent Corpora are becoming the largest parallel corpora worldwide. The Sentence-Aligned European Patent Corpus already contains more than 130 million aligned sentence pairs with very low error rates.

Together with the National Offices of Member States, the EPO aims at further significantly enlarging the POST-GRANT corpus. For non European languages, cooperation is ongoing with the other large IP offices. For these languages the PCT and PRIODOC-A corpora could be considered, potentially also PRIODOC-PRIOTRAN.

Following the London Agreement, and in view of the progress of the Unitary Patent, the number of human made translations will decrease, whereas the need for machine translation of patents will grow. Therefore it is an ideal moment to assemble the corpora now.

In order to make best use of these corpora, an agreement with Google has been signed. The EPO will use Google Translate™ technology to offer translation of patents on its website into 28 European languages, as well as into Chinese, Japanese, Korean and Russian. The EPO will provide Google access to its entire corpus of translated patents to enable Google to optimize its machine translation technology for the specific language used in patent registrations. There is no financial component involved and the agreement is non-exclusive.

References

Dobrinkat, Marcus and Jaakko J. Väyrynen. 2010. Experiments with Domain Adaptation Methods for Statistical MT: From European Parliament Proceedings to Finnish Newspaper Text. *Proceedings of the 14th Finnish Artificial Intelligence Conference STeP 2010, pp. 31–38.*

Gale, William A. and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics 19 (1), pp. 75–102.*

Huet, Stéphane et al. 2010. The RALI Machine Translation System for WMT 2010. *Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, pp. 103-109.*

Koehn, Philipp. 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. *http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl.pdf.*

Lu, Bin et al. 2010. Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. *The 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010). Beijing, China. August, 2010.*

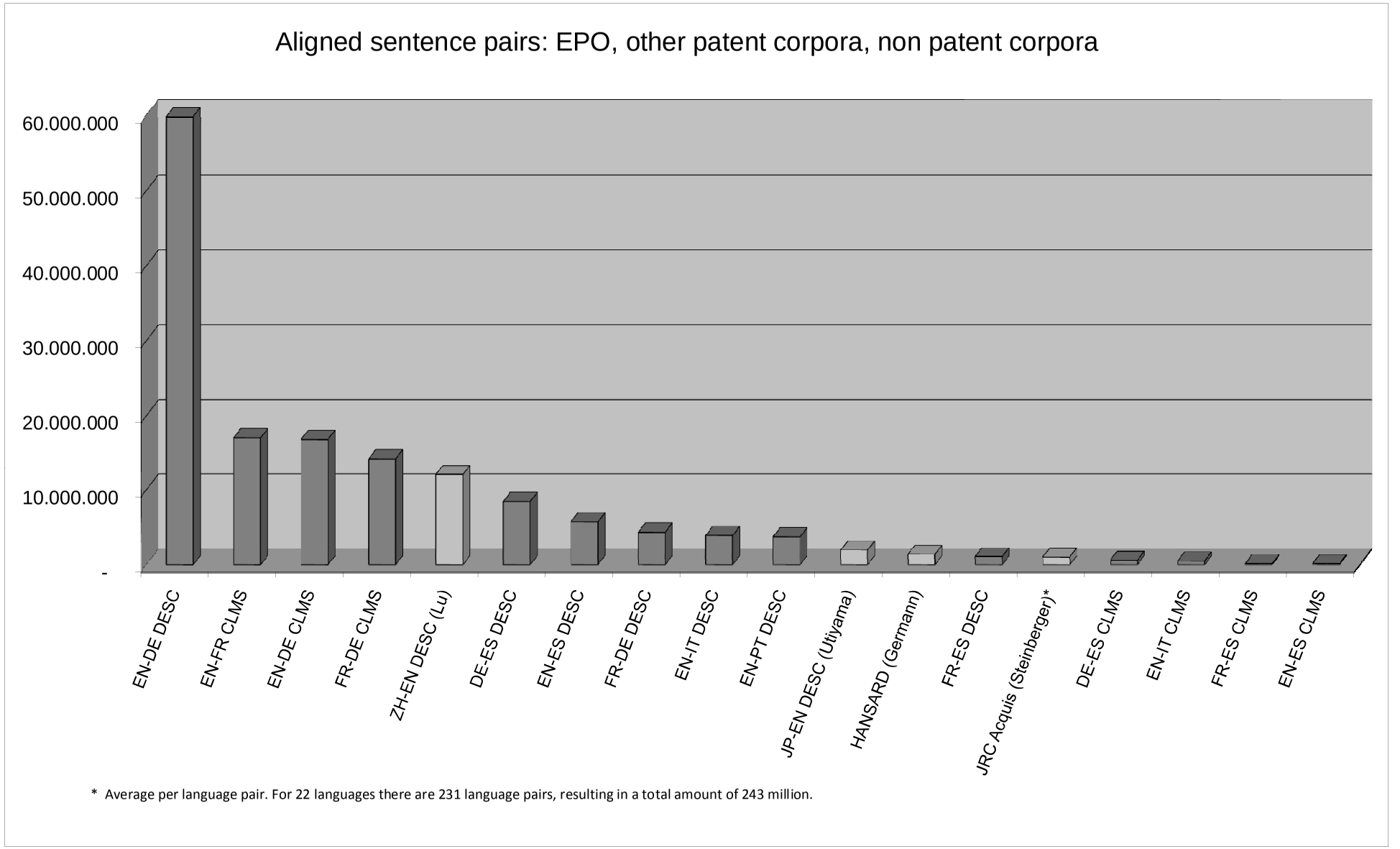
Och, Franz Josef. 2005 Statistical Machine Translation: Foundations and Recent Advances. *Tutorial at MT Summit 2005, Phuket, Thailand.*

Sennrich, Rico and Martin Volk. 2010. MT-based Sentence Alignment for OCR-generated Parallel Texts. *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, CO.*

Tinsley, John, Andy Way and Páraic Sheridan. PLuTO: MT for Online Patent Translation. *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, CO.*

Utiyama, Masao and Hitoshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. *MT Summit XI, pp. 475–482.*

Yasuda, Keiji et al. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. *The Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, pp. 655-66.*


Aligned sentence pairs: EPO, other patent corpora, non patent corpora
60.000.000
50.000.000
40.000.000
30.000.000
20.000.000
10.000.000
-
EN-DE DESC
EN-FR CLMS
EN-DE CLMS
FR-DE CLMS
ZH-EN DESC (Lu)
DE-ES DESC
EN-ES DESC
FR-DE DESC
EN-IT DESC
EN-PT DESC
JP-EN DESC (Utiyama)
HANSARD (Germann)
FR-ES DESC
JRC Acquis (Steinberger)*
DE-ES CLMS
EN-IT CLMS
FR-ES CLMS
EN-ES CLMS
* Average per language pair. For 22 languages there are 231 language pairs, resulting in a total amount of 243 million.