Year: 2018

Parallel Corpora, Terminology Extraction and Machine Translation

Volk, Martin

Parallel Corpora, Terminology Extraction und Machine Translation

Martin Volk
University of Zurich
Institute of Computational Linguistics
volk@cl.uzh.ch

Abstract

In this paper we first give an overview of parallel corpus annotation, alignment and retrieval. We present standard annotation methods such as Part-of-Speech tagging, lemmatization and dependency parsing, but we also introduce language-specific methods, e.g. for dealing with split verbs or truncated compounds in German. We argue for careful sentence and word alignment for parallel corpora. And we explain how word alignment is the basis for a wide range of applications from translation variant ranking to terminology extraction. We conclude with a discussion of the latest developments in Machine Translation.

1 Introduction

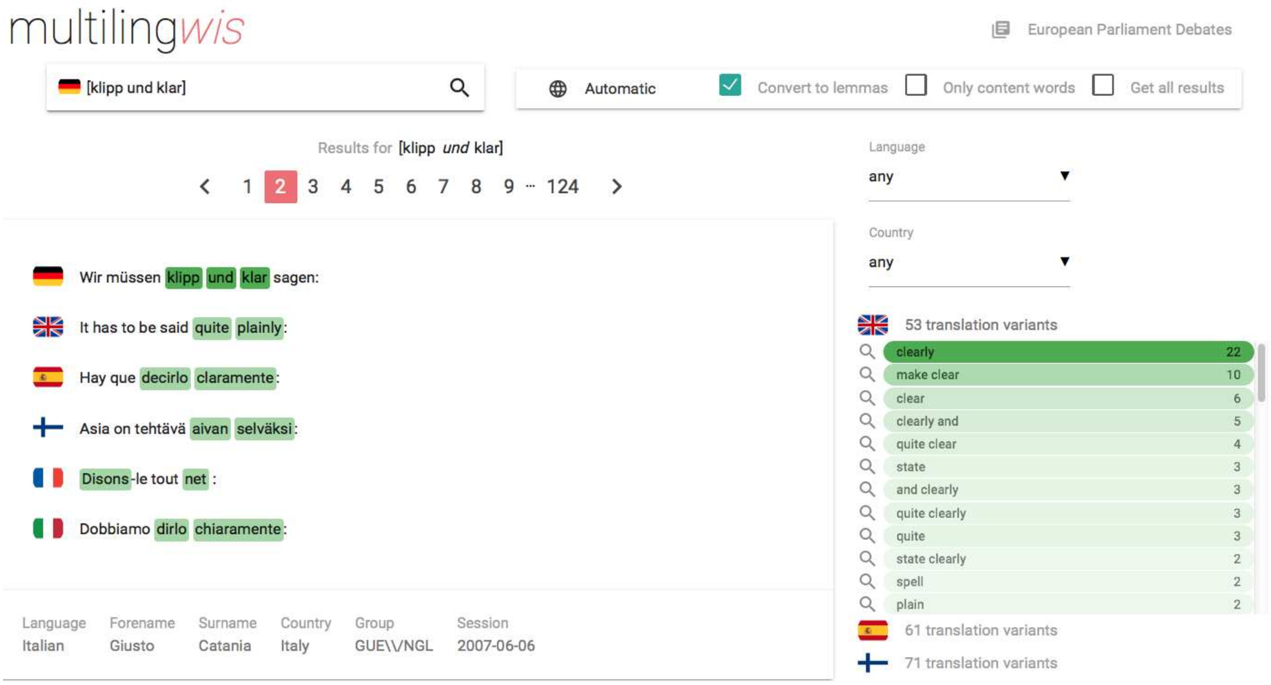
In recent years an increasing number of large parallel corpora have become available for research in natural language processing. The best known is Europarl with the proceedings of the European parliament (Koehn, 2005, Graën et al., 2014) with around 50 million tokens in the languages of the European Union. Other well known multilingual and multiparallel corpora are JRC Acquis (Steinberger et al., 2006) with the EU law collection, OpenSubtitles (Lison and Tiedemann, 2016), United Nations documents (Ziemski et al., 2016), and collections of patent applications (Junczys-Dowmunt et al., 2016b).

Switzerland is a country with four official languages (French, German, Italian and Rumansh) and because of the many international companies and organizations in Switzerland English is becoming ever more popular. Therefore there is a constant need for translations between these languages and this is a natural basis for a plethora of parallel corpora. We have taken advantage of this situation and collected and annotated various Swiss parallel corpora.

Our research group specializes in building parallel corpora for special domains which span over time: We have digitized parallel texts from the Swiss Alpine Club in French and German from 1957 until today (Göhring and Volk, 2011), banking texts in English, French, German and Italian from 1895 up to the present (Volk et al., 2016a), and the announcements of the Swiss federal government (DE: Bundesblatt, FR:Feuille fédérale, IT:Foglio federale) since 1849.

In the current paper we will focus on the latest methods in the automatic annotation and alignment of parallel corpora. We will argue that word alignment across languages improves annotation. We focus on parallel corpora for linguistic and translation studies, but we believe that parallel corpus search systems are also interesting for language learners for viewing translation variants in context and for terminologists who want to extract terms or verify domain-specific language usage.

The paper is structured as follows. In section 2 we describe corpus annotation methods. Section 3 is devoted to alignment techniques and their benefits for corpus annotation such as word sense disambiguation with practical applications in lemma disambiguation and named entity recognition. In section 4 we give usage examples of our parallel corpora including translation discovery and translation error detection. Section 5 describes the latest developments in using parallel corpora for machine translation.



Wir müssen klipp und klar sagen:

It has to be said quite plainly:

Hay que decirlo claramente:

Asia on tehtävä aivan selväksi:

Disons-le tout net :

Dobbiamo dirlo chiaramente:

Language	Forename	Surname	Country	Group	Session
Italian	Giusto	Catania	Italy	GUE\\/NGL	2007-06-06

53 translation variants

clearly	22
make clear	10
clear	6
clearly and	5
quite clear	4
state	3
and clearly	3
quite clearly	3
quite	3
state clearly	2
spell	2
plain	2

61 translation variants

71 translation variants

Figure 1: A Multilingwis query with hits in six Europarl languages

2 Corpus Annotations

Corpus building starts with corpus collection, cleaning and tokenization. The latter is often language-specific and therefore multilingual corpora require language identification. Typically identification is done on the sentence level. For each sentence we compute the language in order to be able to use the appropriate processing tools during corpus annotation. Using, for instance, a Part-of-Speech tagger, that was trained for one language, for the annotation of a sentence in another language will give erratic results. Therefore language identification on the sentence level is of paramount importance for all texts with mixed languages.

2.1 General Corpus Annotation

Part-of-Speech tagging is standard procedure when the corpora are meant for linguistic research. There are a number of PoS taggers available with parameter files for many large languages of Europe. Most of the time the parameter files are the result of training the taggers on newspaper texts. This means that the taggers work best on newspaper texts and gradually worse the more the corpus material differs from newspapers.

Often PoS tagging also provides lemmas. For example, the TreeTagger outputs lemmas for word form - lemma pairs that it has seen in its training corpus. For all other word forms the corpus builder may provide a tagger lexicon with additional pairs. These pairs lack the probabilities that tagger training derives from a manually annotated corpus, but for word forms with little or no PoS ambiguity the extension of the tagger lexicon is still useful. So, any additional lemma information from other corpora or from dictionaries or morphological analyzers are valuable.

Recent parallel corpora have been annotated with more annotation layers: Named entity recognition (NER) is a popular method for a first step towards semantic annotation. Typically it involves the recognition of person names, location names and organization names which are the central classes when processing newspaper texts. Special text types may require other name classes (e.g. event names) or more fine-grained distinctions. For example, in our parallel corpus of Alpine texts we sub-classify toponyms into the name classes of mountains, glaciers, lakes, valleys, cabins, and cities. Toponyms are essential for the mountaineering reports and therefore very frequent.

Shallow NER includes only the recognition and classification of names. A deeper analysis includes co-reference resolution (Ebling et al., 2011) and entity linking (sometimes called grounding). Monolingual co-reference resolution will deal with mention variants like *Grand Combin = Combin*, while multilingual co-references will catch translation variants as e.g. *DE:Matterhorn = FR:Combin = IT:Combino*. Monolingual co-references might include anaphora resolution and will thus allow for investigating coherence phenomena in texts.

Lately, dependency parsing has become available for many languages (e.g. Maltparser, Spacy, Stanford, ...). These parsers allow for the efficient analysis of large corpora with a labeled attachment score of 80-90% (McDonald and Nivre, 2011) and higher values for unlabeled attachment. Even though parsing is far from perfect, the automatically assigned syntax information opens a whole new chapter for corpus studies. For example, searches for verb-object relations no longer need to speculate on co-occurrence in some arbitrary range, but can be conditioned on parsing evidence. In this way, we find candidates for support verb constructions like *to take into consideration* or for verb sub-categorizations for particular prepositional objects like *to wait for*.

2.2 Language-specific Corpus Annotation

In addition to these general annotation considerations, many languages will have specific requirements. Compounding languages like German or Swedish will profit from compound segmentation in lemmatization. E.g. a German text might mention the compounds *Montblanc-Besteigung, Mont-Blanc-Expedition, Montblancgipfel* with or without a hyphen, and we will miss the mountain name *Montblanc* if we do not split these compounds and normalize the spelling variants. Splitting and normalization also facilitate cross-lingual word alignment since it reduces 1-to-many alignments.

Another example of a language-specific annotation is the re-attachment of German verb prefixes that occur separated in the sentence. In example sentence 1 the prefix *auf* (EN: on) needs to be re-attached to the verb stem *fällt* (EN: falls) in order to compute the correct verb lemma *auffallen* (EN: to strike, to notice) (Volk et al., 2016b).

(1) Selber **fällt** mir der kleine Fehler aber kaum **auf**.
EN: However I do not notice the little mistake.

Our re-attachment algorithm is based on PoS tags and re-attaches the separated prefix to the most recent preceding finite verb form when this results in a valid German prefix verb (from a manually curated list of about 8000 such verbs). It works with 96.8% precision when evaluated against manually re-attached prefixes in the TüBa/DZ treebank.

2.3 Exploiting Parallel Corpora for Annotation

Traditionally most corpus annotation is done monolingually. This means that PoS tagging, lemmatization and parsing of e.g. a German corpus is done irrespective of a parallel text in English or any other language. However the parallel text may help to disambiguate and thus to improve the annotation precision on many levels. Most obviously the parallel corpus may help to determine the correct word sense in a given sentence. For example, the word *Mönch* in our Alpine corpus may refer to a prominent mountain in central Switzerland or to a *monk* (= male person in a monastery). If the corresponding sentence in the English or French translation also contains the word *Mönch*, then it is clear that the ambiguous word in the German sentence refers to the mountain name.

We have developed a similar kind of disambiguation for lemmas. For example, the German word form *gehört* may have the lemma *hören* (EN: to hear) or *gehören* (EN: to belong). Depending on the corresponding sentence in English or any other language we can easily compute the correct lemma for the ambiguous word in the German sentence (Volk et al., 2016a). Of course, this kind of knowledge transfer between the languages in a parallel corpus presupposes word alignments across the languages.

3 Aligning Parallel Corpora

Document alignment is the starting point of all alignment activities. If a corpus is built on OCRed text or on web-crawled text, then document alignment requires article boundary detection and subsequent document alignment based on properties such as author names, article lengths and publication dates.

The next step is sentence alignment which is a precondition for any exploitation of parallel corpora. Sentence alignment can be computed efficiently based on length comparisons (based on character counts), co-occurring numbers, names and cognates. Hun-Align is a popular sentence aligner that implements a two-pass algorithm which does a rough alignment and builds a bilingual dictionary in the first pass and uses this dictionary in the second pass for refined sentence alignment. It works well for parallel texts that have corresponding sentences in the same order (monotonicity condition) and with few omissions.

Finally we compute word alignments through GIZA, the Berkeley Aligner or FastAlign. Word alignment finds corresponding words or phrases in aligned sentence pairs. It can be computed on word forms or lemmas. Word alignment enables e.g. sorting the hits after translation variants (which may correspond to different word senses). It also enables annotation transfer (e.g. transferring name classes across languages) and cross-language disambiguation. Word alignment has opened many new avenues for linguistic research and translation studies over parallel corpora.

4 Retrieval from Parallel Corpora

The Corpus Query Workbench has become a defacto standard for retrieval from annotated monolingual corpora. It allows simple and complex queries over words and their associated features (like PoS tags, lemmas, name classes etc.). There is no such standard retrieval tool for parallel corpora.

Different commercial web sites offer searches over parallel corpora as a substitute or complement to bilingual dictionary searches. Most notably these are Linguee, Glosbe and Tradooit (Volk et al., 2014). But on these sites the texts do not have any linguistic annotation.

SketchEngine is one of the few search systems that allows for query conditions to be specified over both sentences in a parallel sentence aligned corpus. But it does not exploit automatically computed word alignment.

We are working towards such a flexible and powerful retrieval tool for parallel corpora. Our prototype system, Multilingwis[1], allows for word form or lemma searches, for fixed sequence and bag-of-word searches and for the exclusion of function words from queries. We have also included the option to filter for source language. In this case, the query (in any of the supported languages) results only in hits with utterances that originate in a specific source language. For example, I could search for German *Binnenmarkt* only in those cases where the original utterance is in French. This allows the researcher to distinguish between searches in original texts versus translations.

In a second prototype we have experimented with database searches over multi-parallel texts including displays of the hits as parallel dependency trees with PoS tags and word alignment (see figure 2).

Word alignment provides the basis for many application scenarios. For example, word aligned corpora allow for translation discovery. We built a parallel corpus English-German of film and TV subtitles. When we queried for the German word *fragen* we found the obvious English translation variants *to ask, to say, to wonder, to question* (in this order of frequency). But next in the ranked list we found *to go* which looked like an alignment error at first sight. But on closer inspection we discovered that this is a real translation option for German *fragen* as in the following example sentences.

(2) DE: ... und sie **fragte** “Was ist das?”.
EN: ... and she **goes**, “What’s that?”

(3) DE: Ich war jung. Ich **fragte** “Wo ist England?”.
EN: I was young. I **went**, “Where’s England?”

A special case of translation discovery is translation error detection. For example, we checked for translation variants of month names. When we queried for English *July* we found that in about 1% of the translations it is erroneously translated with German *Juni*, French *Juin*, and Spanish *Junio* all

[1] pub.cl.uzh.ch/purl/multilingwis

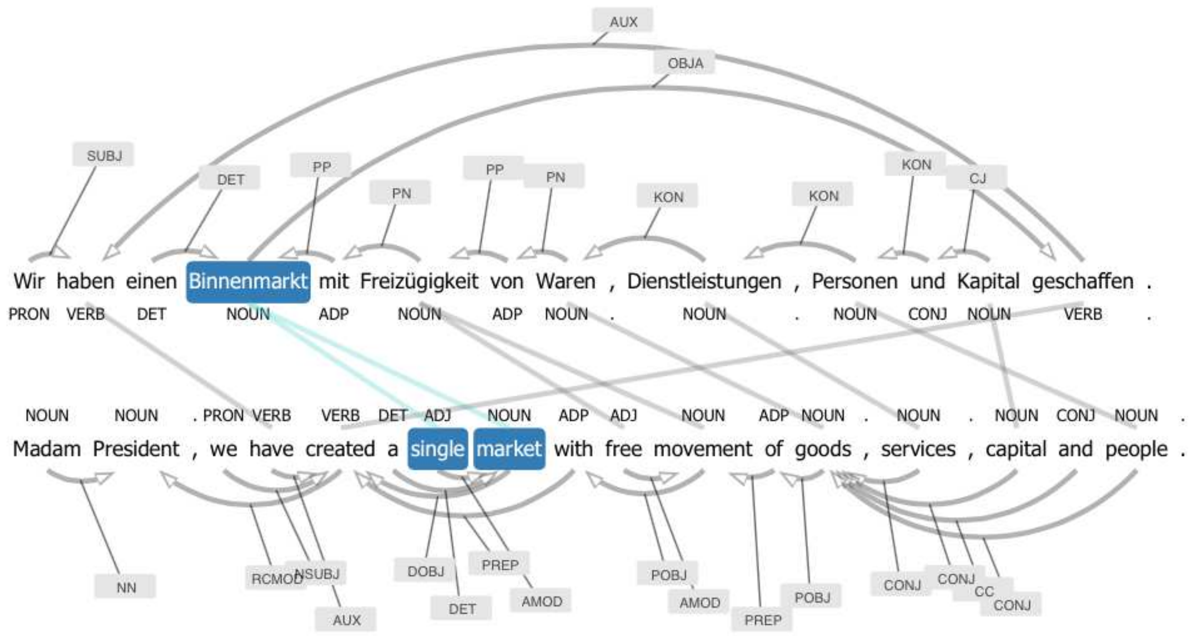

AUX
OBJA
SUBJ
DET
PP
PN
PP
PN
KON
KON
KON
CJ
Wir haben einen Binnenmarkt mit Freizügigkeit von Waren , Dienstleistungen , Personen und Kapital geschaffen .
PRON VERB DET NOUN ADP NOUN ADP NOUN . NOUN . NOUN CONJ NOUN VERB .
NOUN NOUN . PRON VERB VERB DET ADJ NOUN ADP ADJ NOUN ADP NOUN . NOUN . NOUN CONJ NOUN .
Madam President , we have created a single market with free movement of goods , services , capital and people .
NN
RCMOD
NSUBJ
AUX
DOBJ
DET
PREP
AMOD
POBJ
AMOD
PREP
POBJ
CONJ
CONJ
CC
CONJ

Figure 2: Query result for “single market” on Europarl German-English with full annotation display

meaning *June*. We observed this confusion in many of our parallel corpora. Obviously the similarity of the month names *June* and *July* is disturbing for the human translators.

Another application of parallel corpora is synonym detection through mirroring. Starting from a query in one language and then querying back from a translation hit in another language will lead to synonyms in the starting language. For example, when we searched for the idiomatic German adverb phrase *klipp und klar*, we found that *quite clear* and *very clearly* are among the top English translation variants in Europarl (cf. figure 1). Now, if we query for *very clearly* in the opposite direction, we get German *sehr deutlich, ganz klar* and *ganz eindeutig* which are synonymous expressions for the initial query phrase *klipp und klar*.

Queries over parallel corpora provide translation variant ranking based on corpus frequencies. These frequencies are obviously dependent on the corpus (more precisely on the textual domain). For example when querying for the German word *Leiter* (EN: leader or ladder or electrical conductor) we get different rankings for the French translation variants in our Alpine corpus in comparison to our corpus of Swiss laws. In the Alpine corpus the French *échelle* (EN: ladder) is ranked second after *chef* whereas in the Swiss Laws in French *échelle* is only on fifth place after *conducteur, directeur, chef, responsable*.

5 Machine Translation

Machine Translation research dates back to the 1950s. The history has been told in many publications and can therefore be cut short here.

5.1 Old-style Machine Translation

The first approach to MT was based on manually crafted rules for the automatic analysis of the source sentence, for the transfer to the target language and for the generation of the output sentence on the target side. This was a cumbersome and time-consuming process that relied on large bilingual dictionaries and complex grammar rules. Some rule-based MT systems achieved good translation quality for limited topical domains. The most severe limitation of rule-based MT was the lack of a good ranking between possible translation variants and the huge effort needed for adapting an MT system from one domain to another.

These issues were partially remedied through Statistical Machine Translation (SMT) which became prominent in the 1990s and reached its breakthrough a decade later. SMT learns translations from large parallel corpora. Bilingual dictionaries and grammar rules are no longer needed. When a parallel corpus of sufficient size and quality is available and aligned, the SMT system can be built in a matter of days.

In such a way Google was able to build up Google Translate from few languages in 2005 to more than 100 languages to date. Others have built domain-specific systems that outperformed Google Translate in specific topical domains and genres such as TV subtitles. In recent years pure SMT has hit a quality ceiling in particular when translating into morphologically rich languages and between languages with distinctly differing word order.

5.2 Neural Machine Translation

Neural Machine Translation has achieved its breakthrough at the WMT Conference 2016 in Berlin when Edinburgh University's NMT systems ranked first for many language pairs (Sennrich et al., 2016a). In the same year Google published a a seminal paper (Wu et al., 2016) demonstrating a jump in MT quality with the introduction of Neural MT.

Machine learning based on neural networks has been discussed for many years. However progress was limited since it is computationally expensive and thus requires powerful hardware. With the advent of modern GPUs the use of neural networks has become wide-spread and is now known as Deep Learning.

Neural Machine Translation is based on a so-called *encoder-decoder network*. It consists of two recurrent neural networks, one is an encoder which reads the source sentence and produces a sequence of hidden states, and the second is a decoder which predicts each word in the target sentence based on the previously produced target words and a source context, either the last hidden state of the encoder, or a weighted average of the source states in *attentional* models (Bahdanau et al., 2015).

A central notion for NMT are word embeddings, i.e. a mechanism to convert words into numbers so that the neural network can process them. Through word embeddings semantically similar words get similar numerical values. For example, *house* and *building* will be positioned close to each other in the meaning space, whereas *house* and *cloud* will be further apart.

NMT operates with a fixed vocabulary. But real-world translation has to deal with new words constantly. To-date the most successful solution to this problem is to translate via subword units (Sennrich et al., 2016b). The intuition is that various types of words can be translated via smaller units than words, e.g. names can be translated via transliteration, compounds can be handled via splitting and translating the parts, and cognates and loanwords can be processed via phonological and morphological transformations. The subword method reduces the number of out-of-vocabulary words considerably and is currently used in most NMT systems.

As a result NMT output is much more fluent than SMT output. NMT is essentially a powerful language model (on the target language) which is triggered by the source language. Moreover NMT captures more context in the source sentence. The translation choice of every word in the source sentence is conditioned on all other words in the sentence (whereas in SMT this conditioning was limited to adjacent phrase pairs).

Still, many MT problems persist even in NMT. How can we ensure terminology-consistent translation? Many large companies and organizations have compiled terminology databases to ensure the un-ambiguous and consistent translation of technical terms. Chatterjee et al. (2017) have suggested an approach to guide NMT decoding based on external terminology lists which works for words that are known to the NMT system and unknown ones.

Another concern are missing words in the translation. Currently there is no guarantee that the NMT output in the target language has translated all pieces of information from the source. Sometimes there are omissions of modifiers and even negation particles that change the meaning of the output sentence considerably. Koehn and Knowles (2017) discuss more challenges in NMT including larger amounts of training texts needed, difficulties in translating highly inflected words, and worse translation results on very long sentences.

Prominent NMT applications are Google Translate[2], Systran Pure NMT[3], and DeepL[4]. In particular the latter - though currently only available for 7 languages - has attracted much attention because of its high translation quality.

In the language industry, the increased fluency of NMT output has been greeted as a welcome improvement (Junczys-Dowmunt et al., 2016a). However, it is also acknowledged that this requires even more attention and a higher cognitive workload for professional translators who are post-editing MT output. The errors of the MT system are now even more difficult to spot.

6 Conclusion

We have outlined a number of issues for the annotation and alignment of parallel corpora. We have argued that there are standard corpus annotation tools that work for many languages, and there are language-specific annotations that, for example, deal with separated verb prefixes in German. In addition, parallel corpora require automatic alignment not only on the sentence level but also for words and phrases. This level of alignment opens for many new applications such as translation discovery and term extraction.

Automatic word alignment is the off-spring of work on the development of statistical machine translation systems. SMT has been the dominant paradigm in MT research for the last 25 years. Recently it has been superseded by neural machine translation which produces better and in particular more fluent translations.

Acknowledgments

We would like to thank Johannes Graën for preparing the multiparallel Europarl corpus and Christof Bless for the visualizations of parallel parsed sentences. This research was supported by the Swiss National Science Foundation under grant 105215_146781 for "SPARCLING: Large Scale PARallel Corpora for LINGuistic Investigation" (2013-2017) a joint project with Marianne Hundt and Elena Callegaro at the English Department of the University of Zurich.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of 3rd International Conference on Learning Representations (ICLR)*, San Diego.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Conference on Machine Translation (WMT)*, pages 157–168, Copenhagen.

Josep Maria Crego et al. 2016. Systran's pure neural machine translation systems. In *arXiv:1610.05540.*

Sarah Ebling, Rico Sennrich, David Klaper, and Martin Volk. 2011. Digging for names in the mountains: Combined person name recognition and reference resolution for German alpine texts. In *Proceedings of The 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan.

Anne Göhring and Martin Volk. 2011. The Text+Berg corpus: An alpine French-German parallel resource. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier.

Johannes Graën, Dolores Batinic, and Martin Volk. 2014. Cleaning the Europarl corpus for linguistic applications. In *Proceedings of KONVENS*, Hildesheim.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016a. Is neural machine translation ready for deployment? A case study on 30 translation directions. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, December.

Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe Mazenc. 2016b. Coppa v2.0: Corpus of parallel patent applications building large parallel corpora with gnu make. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora at LREC*, Portorož, Slovenia, May.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of First Workshop on Neural Machine Translation*, Vancouver.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket.

[2] https://translate.google.com; At the time of writing about half of the languages supported by Google Translate are handled with NMT and the rest still with SMT.

[3] https://demo-pnmt.systran.net; see also (Crego et al., 2016)

[4] https://www.deepl.com

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, Berlin. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *In Proceedings of the 54th Annual Meeting of the ACL*, pages 1715–1725, Berlin.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Carmelia Ignat, Tomaz Erjavec, Dan Tufiş, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, Genoa.

Martin Volk, Johannes Graën, and Elena Callegaro. 2014. Innovations in parallel corpus search tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016a. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of KONVENS*, Bochum.

Martin Volk, Simon Clematide, Johannes Graën, and Phillip Ströbel. 2016b. Bi-particle adverbs, PoS-tagging and the recognition of German separable prefix verbs. In *Proceedings of KONVENS*, Bochum.

Yonghui Wu et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv e-prints*, September.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.