Parallel Data, Tools and Interfaces in OPUS

Jörg Tiedemann

Department of Linguistics and Philology
Uppsala University, Uppsala/Sweden
jorg.tiedemann@lingfil.uu.se

Abstract

This paper presents the current status of OPUS, a growing language resource of parallel corpora and related tools. The focus in OPUS is to provide freely available data sets in various formats together with basic annotation to be useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. In this paper, we report about new data sets and their features, additional annotation tools and models provided from the website and essential interfaces and on-line services included in the project.

Keywords: parallel corpora, bitext, alignment

1. Introduction

OPUS is a growing resource of freely accessible parallel corpora. It also provides tools for processing parallel and monolingual data as well as several interfaces for searching the data, which makes it a unique resource for various research activities. Parallel corpora are, for example, essential for most data-driven approaches to machine translation. Statistical machine translation is one of them being in constant need of more training data. But this is not the only application area. For example, people working in translation studies have discovered the multilingual search interfaces provided by OPUS for empirical studies on translation data. Parallel corpora have also been used for word sense disambiguation, paraphrasing and the extraction of monolingual lexico-semantic information to name a few other applications.

Currently, parallel corpora in sufficient data sizes are only available for a few language pairs and these sources usually cover only some specialized domains. With OPUS we try to improve the situation by compiling additional data sets on a large scale in order to provide data for many other, often under-resourced languages and domains. The overall goal of the OPUS project is to make parallel resources freely available, especially emphasizing the support of low density languages.

In this paper, we report recent developments related to large amounts of new data sources, additional tools and on-line interfaces. The following section is devoted to data resources. Thereafter, we discuss provided tools and, finally, we present search interfaces and plans for future work.

2. Growing Parallel Data Resources

OPUS is probably the largest collection of freely available parallel corpora. The corpus covers over 90 languages[1] and includes data from several domains. Altogether, there are over 3,800 language pairs with sentence-aligned data comprising a total of over 40 billion tokens in 2.7 billion parallel units (aligned sentences and sentence fragments). Figure 1 illustrates the size of the top 100 language pairs included in OPUS. It shows that those sub-corpora are well above 100 million words, which is definitely a considerable size even for data-intensive NLP. The language pair with the largest amount of parallel data is Spanish-English with about 36 million parallel sentences containing roughly 500 million tokens. Even though many of these top-ranked language pairs represent traditionally well-supported languages, there are also various pairs of otherwise poorly resourced languages among the top 100. For example, there are parallel corpora for Romanian-Turkish and Bulgarian-Hungarian with over 100 million words in our collection, which is rarely to be found elsewhere.

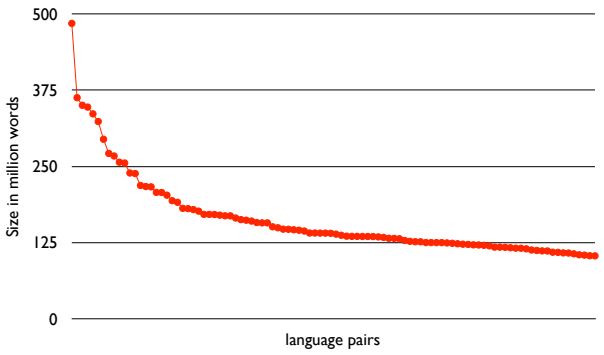


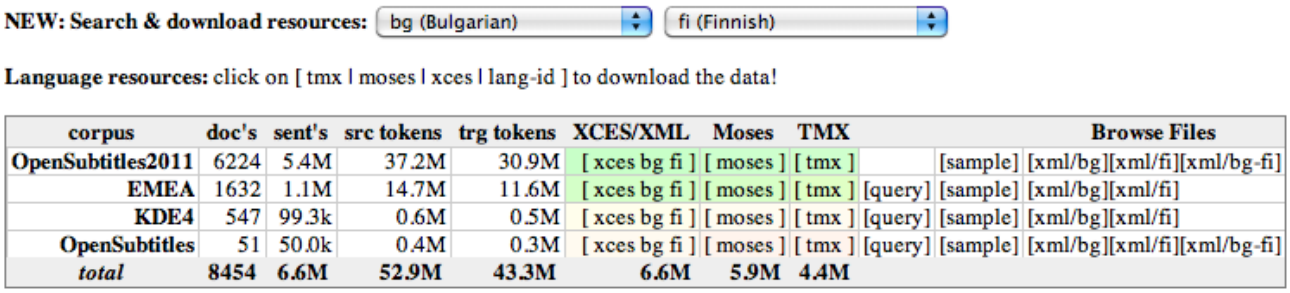
Figure 1: Size of the top-100 language pairs in OPUS.

The largest domains covered by OPUS are legislative and administrative texts (mostly from the European Union and associated institutions), translated movie subtitles and localization data from open-source software projects. There is also a substantial amount of newspaper texts and some other smaller collections from various on-line sources. Some of the resources have previously been presented (Tiedemann, 2009).

Recently, OPUS has been extended by several large collections. The following gives a brief overview over the major extensions:

ECB: This corpus contains data from the European Central Bank, which has originally been compiled by Al-

[1] We do not give exact numbers here as it is sometimes debatable if certain languages should be counted as a separate language or not, for example, Brazilian Portuguese.



corpus	doc's	sent's	src tokens	trg tokens	XCES/XML	Moses	TMX			Browse Files
OpenSubtitles2011	6224	5.4M	37.2M	30.9M	[xces bg fi]	[moses]	[tmx]		[sample]	[xml/bg][xml/fi][xml/bg-fi]
EMEA	1632	1.1M	14.7M	11.6M	[xces bg fi]	[moses]	[tmx]	[query]	[sample]	[xml/bg][xml/fi]
KDE4	547	99.3k	0.6M	0.5M	[xces bg fi]	[moses]	[tmx]	[query]	[sample]	[xml/bg][xml/fi]
OpenSubtitles	51	50.0k	0.4M	0.3M	[xces bg fi]	[moses]	[tmx]	[query]	[sample]	[xml/bg][xml/fi][xml/bg-fi]
total	8454	6.6M	52.9M	43.3M	6.6M	5.9M	4.4M			

Figure 2: Searching for parallel data. The example shows a query for Bulgarian - Finnish parallel data. Data resources can immediately be downloaded from the listed results of the query by clicking on the links of the corresponding data formats (XCES/XML, Moses, TMX).

berto Simões (http://alfarrabio.di.uminho.pt/∼albie/). It contains over 30 million parallel translation units in 19 European languages.

MBS: This corpus is a collection of translated newspaper texts from the "Moniteur Belge/Belgisch Staatsblad" compiled by Tom Vanallemeersch (Vanallemeersch, 2010). It contains French and Dutch texts of over 10 million parallel units.

OpenSubtitles2011: This is a major extension of the old subtitle corpus and constitutes now the biggest parallel corpus in OPUS. It contains data from over 50 languages with altogether over 8 billion words.

TEP: This is another subtitle corpus for Persian and English compiled by (Pilevar et al., 2011). This corpus is cleaned and pre-processed with special care for the Persian part and contains about 1.2 million parallel sentences.

WikiSource: This collection is the result of a pilot study of extracting parallel resources from the public domain content collected at WikiSource. Currently, it only contains a Swedish and an English translation of the Bible.

The *OpenSubtitles2011* corpus deserves some special attention. Unique for this collection is not only its size but also its contents. Subtitles cover various genres and time periods and combine features from spoken language corpora and narrative texts including many dialogs, idiomatic expressions, dialectal expressions and slang. It, therefore, represents a quite exceptional resource especially considering its size and language variety. The collection comes from a free on-line resource of user uploads (www.opensubtitles.org). Due to its open nature, some special treatment was required when compiling the corpus. First of all, the original collection contains many duplicates, i.e., subtitle variants of the same movies. Furthermore, it contains a lot of noise in terms of corrupted files and incorrectly tagged files (wrong language, wrong encoding, wrong movie, etc).
Fortunately, the source is sorted by movie and release year which simplifies the matching of subtitles substantially. However, TV series with many episodes are usually tagged with the same identifier and, therefore, are not easily matched with their translations into other languages. Fortunately, we could rely on very fast time-based sentence alignment (Tiedemann, 2007), which enabled us to perform a brute-force search for matching movies among all combinations of subtitles tagged with the same movie ID. For selecting the most appropriate match, we defined the following criteria:

- All subtitles are aligned to one and only one translation in another language.
- At least 50% of the time specified in subtitle time frames has to overlap.
- The best candidate is the subtitle pair with the largest time overlap.
- The ratio between non-empty alignments and empty alignments has to be larger than two.

Using these criteria we could largely filter out non-matching documents. To further clean the data, we also applied automatic detection of language-dependent character encoding using *chared* (Pomikálek and Suchomel, 2011) and automatic language verification using *textcat* (van Noord, 2010). For the latter we trained appropriate language models for Unicode UTF-8 texts for all languages involved in the corpus. These models are also released on our website (http://opus.lingfil.uu.se/trac/wiki/DownloadTools).
However, the actual alignment based on time information is still not perfect due to synchronization differences. Therefore, we ran a second alignment for all subtitle pairs identified in the first run using lexical synchronization as proposed in (Tiedemann, 2008). For this it was necessary to create bilingual dictionaries for all language pairs involved. This was done by running automatic word alignment on the entire parallel data set created in the previous step. We used GIZA++ (Och and Ney, 2003) and the symmetrization heuristics (grow-diag-final-and) implemented in Moses (Koehn et al., 2007) to extract probabilistic phrase tables used in statistical machine translation. These word alignments are also freely available from OPUS (http://opus.lingfil.uu.se/OpenSubtitles2011/wordalign/).
From the phrase translation tables we can now extract highly reliable lexical translations even though they are based on the alignment of partially noisy corpora. We

applied heavy filtering using probability thresholds, frequency thresholds and string patterns. In particular, we extracted one-to-one word alignments of words containing at least three, exclusively alphabetic characters that occurred at least twice in the corpus and obtained conditional phrase translation probabilities $\phi(f|e)$ and $\phi(e|f)$ of equal or more than 0.1. We did not spend much time optimizing these parameters but for most language pairs, this procedure gave us a decent amount of reliable word translations that could be used to find lexical matches in subtitle pairs. The dictionary-based synchronization techniques presented in (Tiedemann, 2008) were then used to re-align all subtitle pairs. These improved sentence alignments can now be found together with the bilingual dictionaries used for synchronization at the following URL: http://opus.lingfil.uu.se/OpenSubtitles2011/srtalign/. A final alternative provided for the OpenSubtitles2011 corpus is an alignment based on hunalign (Varga et al., 2005): http://opus.lingfil.uu.se/OpenSubtitles2011/hunalign/.

3. Resource Availability

Another improvement of recent versions of OPUS is the availability of various download formats for all subcorpora. We now provide all data in their native XML format (using the XCES Align DTD for sentence alignment), in Translation Memory eXchange format (TMX) and in plain text format (for Moses/GIZA++). We also integrated a special interface for searching the entire collection for specific language resources. An example is given in figure 2. Furthermore, we also started a Wiki (http://opus.lingfil.uu.se/trac) with further information about the corpus. Moreover, the website and OPUS related data are now stored on a dedicated server to reduce interference with other processes (http://opus.lingfil.uu.se) and users.

4. Annotation Tools

Another goal of OPUS is to provide tools for processing parallel and monolingual data. The multi-lingual nature of the corpus makes it necessary to process its documents in language-specific ways. It is still on-going work to collect dedicated processing pipelines for all the languages included in OPUS. Many of them are still processed with generic fall-back approaches. However, we started collecting tools for many language improving the annotation of the corpus. These tools will also be provided for downloading (if license agreements permit) to add yet another value to the project.

One major plan for the future is to add dependency information to data in our collection. For this, we will rely on statistical parsers trained on available treebanks. This also presupposes part-of-speech tagging which is already done for some languages and parts of the corpus. Our pipeline will be based on state-of-the-art toolboxes such as *hunpos* (Halácsy et al., 2007) and *MaltParser* (Nivre et al., 2007) and pre-trained models for various languages. These models are available from the OPUS Wiki (http://opus.lingfil.uu.se/trac/wiki/DownloadTools) and currently support the following languages (see table 1).

Language	POS-Tagger	Parser
Catalan	SVMTool	malt
Czech	HunPos	malt
Chinese	Zpar	Zpar
Danish	HunPos	malt
Dutch		malt
English	HunPos	malt
French	MElt	malt
German	HunPos	malt
Hungarian	HunPos	
Italian	TextPro	malt
Portuguese	HunPos	malt
Russian	HunPos	malt
Slovene	HunPos	malt
Spanish	SVMTool	malt
Swedish	HunPos	malt

Table 1: Currently available annotation tools. "malt" refers to MaltParser v1.4.1 models, "zpar" refers to a statistical parser with language-specific features for Chinese and English (Zhang and Nivre, 2011), SVMTool (Giménez and Màrquez, 2004) and MElt (Denis and Sagot, 2009) are statistical taggers with language-specific models required for the respective dependency parser models listed above.

Other tools that we frequently use for compiling OPUS corpora are listed at http://opus.lingfil.uu.se/tools.php. They include the time-based sentence aligner for movie subtitles, various scripts for converting and browsing data and other annotation tools used for various languages such as the TreeTagger (Schmid, 1994).

5. Search Interfaces

Beside the actual data sets and tools for processing them, OPUS also provides on-line interfaces for searching its database. We developed multilingual concordance tools built on top of the Corpus Workbench (Evert and Hardie, 2011) for this purpose. Figure 4 shows parts of a simple example query on the multilingual subtitle data.

18355493	Hey , **honey** , we' re coming !
es	Oye , querida , ya vamos !
pt	Querida , estamos a chegar !
sv	Vi kommer nu !
18615253	You okay , **honey** ?
es	¿ Estas bien , cariño ?
pt	Estás bem querida ?
sv	Är du okej , älskling ?
18615685	It' s okay , **honey** .
es	Está bien , cariño .
pt	Está tudo bem , querida .
sv	- Det blir bra , det blir bra , älskling .

Figure 4: Searching for "honey" in English subtitles aligned to Spanish (es), Portuguese (pt) and Swedish (sv).

For some corpora, there are also additional search interfaces that make use of corpus-specific information such as the speaker tags in the Europarl corpus. These interfaces also support additional features using linguistic annotation

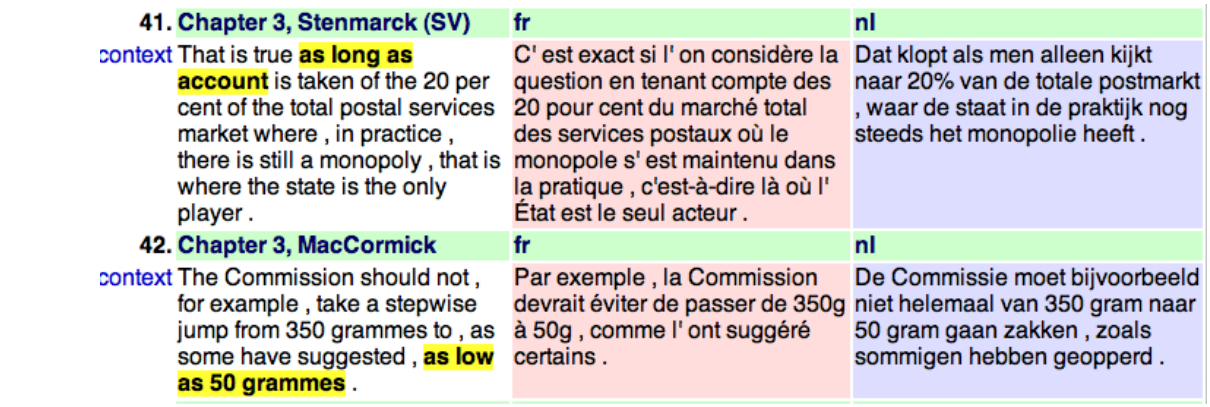


Figure 3: Querying an annotated parallel corpus using a corpus-specific search interface (Europarl). The example is taken from the following query: `"as" [tnt="JJ.*"] "as" <chunk_type="NP"> []+ </chunk>` (the word "as" followed by an adjective based on TnT-tagger labels, followed by "as" and followed by an NP found by the English chunker).

for some languages. For example, it is possible to illustrate the bracketing structure produced by shallow parsing for English in the Europarl data. Available annotation can be queried using the flexible CQP query language of the Corpus Workbench by all concordance tools in OPUS (see figure 3).

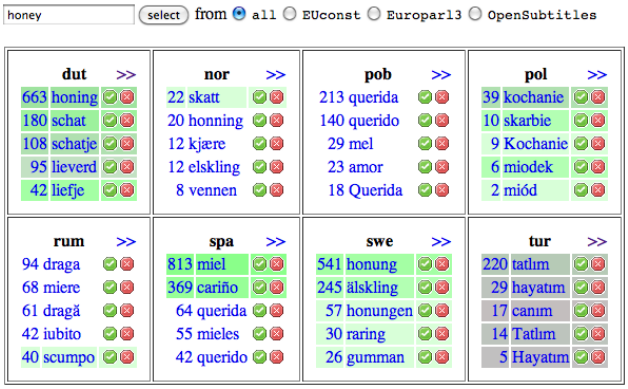


Figure 5: A word alignment database with feedback function. The colors correspond to the proportion between acceptance and rejection of the proposed translation (green = mostly correct, red = mostly incorrect, gray = undecided). Word translations are linked to the concordance tool to obtain real-world examples from OPUS corpora.

Finally, there is also a database of word alignments extracted from a subset of the OPUS corpora. This database represents a rough multilingual dictionary with links to real-world examples. The database entries can be judged by users and can easily be explored through interlinked terms. Figure 5 shows a screenshot of the on-line interface.

6. Conclusion and Future Work

In summary, OPUS presents a unique collection of parallel resources and tools for processing them. A large number of new sources have been added recently and we plan to further extend the corpus in the future. We started already to word-align larger portions of our collection and the results will be available to the public from our website. Furthermore, we plan to add more linguistic annotation. In particular, we like to add dependency information for several corpora. A prototype for the visualization of machine-annotated parallel treebanks is shown in figure 6. Such an interface could also be used to verify and correct parse trees to improve the annotation and possibly to create extended training data.

Another idea is to open the collection to user contributions using automatic upload facilities. Pre-processing and alignment tools could then be run off-line supporting users without technical experience to built new parallel resources on their own. Our idea is to even support the creation of private data collections but the main motivation is to enable external users to contribute to our free collection of parallel corpora.

7. References

P. Denis and B. Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong Kong, China.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. Presentation at Corpus Linguistics 2011, University of Birmingham, UK.

Jesús Giménez and Lluís Màrquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th LREC*.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Poster paper: Hunpos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 209–212, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of*

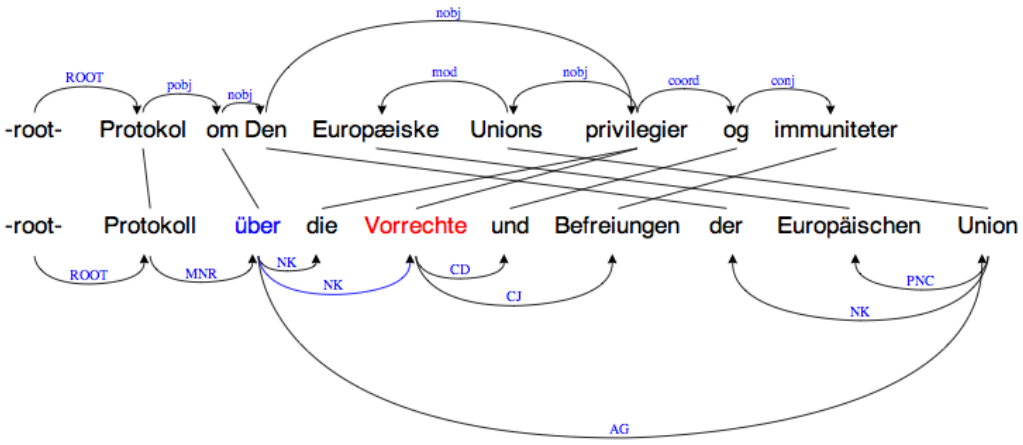


Figure 6: Visualization of aligned parallel dependency trees. A prototypical interface with editing possibilities (tree structure and word alignment).

1023298	want something done ... do it yourself . Come on ,	**honey**	. You can' t die . Wake up ! Where are the stones
3857140	s him . Where are we going , Mama ? It' s a game ,	**honey**	. Like hide and seek ? Yes . Like hide and seek .
4652049	. Is he still alive ? Barely , but don' t worry ,	**honey**	, I think I can save him . You' re a very brave y
5772257	y . Oh ... And ... Aw ! Well , actually , I ate my	**honey**	. - But it made me do it . - Humph . I was asleep
6560977	. Lau . - Yeah ? - What do you think of this one ,	**honey**	? - Please ! It needs a little swag . A scooch of
7687480	where reality is over there , somewhere ... - Oh ,	**honey**	, don' t . and we hide from it over here and pret
7886828	honey , and if you wanted honey , you' d just buy	**honey**	instead of ... apricots . Um , but nevertheless ,
8987992	r downtown . Come on . Here they come ! This way ,	**honey**	. Oh , come on . It? s a shame to hide such a bea
9055714	oesn' t interfere with couples ' bowling , right ,	**honey**	? You ever bowl with Rusty ? It' s a good thing .
9798978	good day . What do you want , Bruiser ? Bruiser ,	**honey**	, come on . We have to go . We' re late . We have
10002428	he school nurse . - It' s against the law . - Oh ,	**honey**	. It' s a girl' s best friend . - A certain kind

Figure 7: A monolingual concordance tool displaying keywords in context.

the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran english-persian parallel corpus. In *CICLing (2)*, pages 68–79.

Jan Pomikálek and Vít Suchomel. 2011. chared – a character encoding detection tool. available from http://code.google.com/p/chared/. website accessed 13 October 2011.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, September. http://www.ims.uni-stuttgart.de/˜schmid/.

Jörg Tiedemann. 2007. Improved sentence alignment for movie subtitles. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*, Borovets, Bulgaria.

Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'2008)*, Marrakesh, Marocco.

Jörg Tiedemann. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Gertjan van Noord. 2010. TextCat - an implementation of a text categorization algorithm. `http://www.let.rug.nl/~vannoord/TextCat/`.

Tom Vanallemeersch. 2010. Belgisch staatsblad corpus: Retrieving french-dutch sentences from official documents. In *Proceedings of LREC 7*, pages 3413 – 3416, Malta.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA, June. Association for Computational Linguistics.