The Text+Berg Corpus
An Alpine French-German Parallel Resource

Anne Göhring, Martin Volk
UZH, Institute of Computational Linguistics
University of Zurich, Switzerland
lastname@cl.uzh.ch

Résumé. Cet article présente un corpus parallèle français-allemand de plus de 4 millions de mots issu de la numérisation d'un corpus alpin multilingue. Ce corpus est une précieuse ressource pour de nombreuses études de linguistique comparée et du patrimoine culturel ainsi que pour le développement d'un système statistique de traduction automatique dans un domaine spécifique. Nous avons annoté un échantillon de ce corpus parallèle et aligné les structures arborées au niveau des mots, des constituants et des phrases. Cet "alpine treebank" est le premier corpus arboré parallèle français-allemand de haute qualité (manuellement contrôlé), de libre accès et dans un domaine et un genre nouveau : le récit d'alpinisme.

Abstract. This article presents a French-German parallel corpus of more than 4 million tokens which we have compiled as part of the digitization of a large multilingual heritage corpus of alpine texts. This corpus is a valuable resource for cultural heritage and cross-linguistic studies as well as for the development of domain-specific machine translation systems. We have turned a small fraction of the parallel corpus into a high-quality parallel treebank with manually checked syntactic annotations and cross-language word and phrase alignments. This alpine treebank is the first freely available French-German parallel treebank. It complements other treebanks with texts in a new domain and genre : mountaineering reports.

Mots-clés : corpus alpin français-allemand, structures arborées parallèles, annotation morphosyntaxique du français.

Keywords: French-German alpine corpus, parallel treebank, French morphosyntactic annotation, Text+Berg, e-Humanities.

1 Introduction

Parallel corpora have become central resources for many areas in natural language processing such as word sense disambiguation, bilingual terminology extraction and machine translation. However most of the large available parallel corpora come from a limited set of domains (parliamentary proceedings, legal texts). We have compiled a sizable parallel corpus of French-German alpine texts. It also differs from previous parallel corpora in that it was built on the basis of printed books that we scanned and OCRized.

Our parallel corpus is a by-product of our effort to digitize and annotate all the yearbooks of the Swiss Alpine Club from 1864 until today. Since 1957 the books have been published in parallel language versions in French and German. We have scanned all books (more than 80,000 pages) until the year 2000. The books from 2001 to 2009 were provided by the Swiss Alpine Club as PDF documents. Overall, this resulted in a parallel corpus of more than 4 million tokens each in French and German.

We selected a small part of this corpus, 1000 sentences from mountaineering reports, to build a parallel treebank. This treebank consists of manually checked syntax structures on both the French and German sentences as well as cross-language word and phrase alignments. There are state-of-the-art automatic tree aligners : the supervised approach reported in (Tiedemann & Kotzé, 2009) outperforms the unsupervised technique described in (Zhechev, 2009). We decided to build a manually checked, high-quality alpine treebank to complete our collection.[1] These treebanks are useful to train and evaluate automatic tree annotation and alignment.

In this paper we first describe the creation of our Text+Berg corpus. We then focus on the French language parts in the mixed language period from 1864 to 1956 and the parallel language period from 1957 to 2009. We give an overview of the number of French articles. But we also look at French sentences scattered throughout the corpus in German articles. This case study of language mixture illustrates the linguistic richness in our alpine corpus. In the final section we present our steps for building the French-German parallel treebank, with particular attention to the annotation of the French treebank.

[1] We have released this treebank as part of our SMULTRON corpus which otherwise consists of parallel treebanks in English, German, Spanish and Swedish for three other text genres. We are distributing the latest version of our multilingual parallel treebank as SMULTRON v 3.0, Volk *et al.* (2010c).

2 Overview of the Text+Berg Corpus

The Text+Berg project aims at the collection of a large corpus of alpine texts to study the evolution of language, culture, technology in this particular domain. We now have digitized the complete series of the yearbooks of the Swiss Alpine Club which contain articles in French, German, Italian and Romansh, all related to one central theme : mountains. Articles comprise club activities, mountaineering, travel and scientific reports on different aspects of the alpine world : climatology, geology, fauna, flora, society, culture, tourism, leisure and sports. There are also collections of bibliographical references, book reviews, and even some poems and lyrics.

As of March 2011 we have scanned and OCR-converted 196 books from 1864 to 2009, corresponding to nearly 87,000 pages with a total of 35.75 million tokens representing 5.6 million types.[2] The result is a multilingual corpus of 9,917 articles in German, 5,998 in French, 162 in Italian, 17 in Romansh, 4 in Swiss-German and even 1 in English : a typical Swiss product indeed. The majority of the texts are untranslated, forming thus a multilingual corpus, but almost 15% is truly parallel, i.e. the same articles appear both in French and German. This parallel subcorpus currently contains 2375 translated articles amounting to 4.7 million tokens in French and 4.2 million tokens in German.

To create this corpus we had to first collect the books, then we cut them open so that we were able to scan them with automatic paper feed. In order to reduce the number of OCR errors we merged the output of two OCR systems (Abbyy FineReader 7 and Omnipage 17) by automatically selecting the respective best output, and applied further automatic OCR corrections patterns (cf. (Volk *et al.*, 2010a)). We manually corrected all tables of content and turned them into tabular format. We finally processed the text, adding structural metadata, e.g. article boundaries, headers, authors, as well as linguistic information like lemmas and part-of-speech tags. We have also annotated geographical entities (mountain, cabin and glacier names) which abound in our corpus.[3]

Many of these steps require different treatment based on the language. We have assigned a language tag to each article (manually) and each sentence (automatically). We used Lingua-Ident[4] to automatically identify the language of each sentence based on character n-grams. This method is not reliable for short character sequences. Therefore sentences under a certain length threshold inherit the language assigned to the article they appear in. In this way we distinguish between English, French, German, Italian and Romansh. In a second step we test for each sentence that was identified as German whether it is written in a Swiss German dialect. This step is based on a word list with typical Swiss German terms that do not have homographs in standard German.

From the first yearbook in 1864 the corpus contains French articles. This is astonishing as the parallel French language yearbook "Echo des Alpes", which was published by the French sections of the Swiss Alpine Club, in turn does not have any German articles. This French row of publication ceased in 1923 which resulted in a remarkable increase of French articles in the German yearbooks. From 1957 the yearbooks have been published in parallel French and German versions, though all the articles were translated only since 1982. The lengths of the articles steadily decreased over time following the typical reading patterns.

average percentage of French articles (sentences)			avg article length (tokens)		
1864-1923	1925-1956	1957-2009	19th	1950s	21th
6.6% (12.7%)	39.3% (37.4%)	45.7% (48.6%)	5974	2613	1150

While the thematic focus on the mountains has stayed the same over the almost 150 year period, the specific topics have changed. In the 19th century the mountaineering descriptions had more of a scientific or discovery tone. The mountains were measured and the routes explored. Over the years mountaineering has turned into a sports activity either for competition or recreation.

2.1 French sentences in German texts

Table 1 shows the language distribution of articles (in rows) and sentences (in columns). For example, the number in column 3 means that our corpus contains 12,392 German sentences within French articles. The numbers are unequally distributed across the six languages, mirroring the global language tendencies of the corpus.

Let us look at the "foreign" sentences in more detail. German and French articles contain sentences in all identifiable languages, but only German sentences appear in articles of all other languages. Italian is under-represented with only 162 articles ; furthermore the

[2]The latest release can be downloaded or accessed online after registration : `http://www.textberg.ch/index.php?id=4&lang=en`

[3]For the annotation steps of our pipeline we used Gertwol for the lemmatization (only for German), TreeTagger for the PoS tagging and an own NER module tailored to the alpine domain (Volk *et al.*, 2010a).

[4]see Lingua-Ident by Michael Piotrowski : `http://search.cpan.org/dist/Lingua-Ident/`

numbers of Italian and Romansh sentences in French and German articles are about the same, though Romansh is marginal in our corpus. The figures for Romansh sentences should be taken with caution as our language identifier sometimes confuses it with other Romance languages. The particular status of the Swiss-German dialect is somehow (is it due to the special language identification procedure, the diglossia, or both ?) reflected in the high proportion of German sentences in the 4 Swiss-German articles and the almost exclusive presence of Swiss-German sentences in German articles (apart from 2 Swiss-German sentences in French articles). English has been identified 1683 times at sentence level but there is only one complete article in English. English sentences appear proportionally as often in German as in French articles.

article language	number of articles	number of sentences						
		German	French	Italian	Romansh	CH-German	English	total
German	9,917	1,166,141	11,607	1481	1490	799	1035	1,182,553
French	5,988	12,392	670,599	1187	1277	2	607	686,064
Italian	162	329	243	15,048	69	0	1	15,690
Romansh	17	12	3	7	771	0	1	794
CH-German	4	28	0	0	0	143	0	171
English	1	4	0	0	0	0	39	43
Total	16,089	1,178,906	682,452	17,723	3,607	944	1683	1,885,315
		62.53%	36.20%	0.94%	0.19%	0.05%	0.09%	100.00%

TAB. 1 – Number of articles and sentences per language in Text+Berg corpus (release 145)

The language statistics reveal the presence of French sentences in articles written (or translated) in German and vice-versa. In the monolingual as well as in the parallel parts of the corpus, we observe a great mixture of languages, mostly German and French, but also Italian, English and Romansh. The general challenge we face is how to find the monolingual units that form this multilingual text landscape.

The source of language mixture are quotations, bibliographical references, direct speech, lyrics, itinerary descriptions and even panel inscriptions. French quotations are often left untranslated in the German articles. In a German article from 1990, the author cites the natural scientist de Saussure directly in French : “La vue que l’on a du haut de l’Etna est sans doute plus étendue et plus riante, mais celle de la chaîne des Alpes, que l’on découvre de la cime du Buet, est peut-être plus étonnante : elle excite dans l’âme une émotion plus profonde, et donne plus à penser au philosophe”. This particularity of our corpus is certainly due to the presupposed language skills of the readers : German and Italian speaking Swiss should be able to understand written French. In other words, like Latin or Greek for classic studies and nowadays English for scientific literature, French is (or was at least supposed to be) intelligible to all readers of the SAC yearbooks and *The Alps*. References offer a second type of multilinguality : either the whole article is a bibliography with entries in different languages or it contains some bibliographical references in the original language(s) they were published in. For example, the French reference “La contribution de l’Alsace à l’exploration des Alpes” has been identified in the German “Bibliographie zur Geschichte und Herstellung alpiner Reliefs, besonders in der Schweiz” (1981).

Sometimes our language identifier fails to recognize a French reference although the character sequence is long enough, like the entry “Baumann, Joseph : Un pionner alsacien de l’alpinisme” from the same German bibliography. Another problematic issue is the inclusion of a French sentence within a German sentence. In this case, although quotation marks and/or colon may separate it explicitely from its surrounding German sentence, the French part is not identified and the individual tokens are implicitely interpreted and PoS tagged as if they were German : “In höhere Regionen führt uns die "Traversée des arêtes des Grandes-Rousses de l’ Etendard au Pic Bayle" ”.

There are many sources of errors, false positive or false negative examples of language mixture, the main cause being the erroneous language setting of the article : it produces many follow up errors given that the language of too short a sentence is assigned to it by default. Similarly, if we fail to identify an article boundary, the whole text up to the next boundary remains under the same language default. Unfortunately, articles are also sometimes merged with advertising text. The presence of many untranslatable named entities increase the probability of their original language, misleading the statistical language identification procedure to wrongly identify a French sentence in German text : “Erwähnt seien etwa Voie des Dames (5b), La pénible (6b), La balade du petit Jules (6b) oder Tranquille champagne (7b+)”. These errors may bias the interpretation of the multilingual characteristics of our corpus.

3 Our French-German Parallel Alpine Treebank

Text Selection Most treebanks are built on newspaper texts. We are convinced that this monoculture is limiting both the applicability of the treebanks and the linguistic research. In our own treebanking we follow a different approach by diversifying over text genres and domains. In the Text+Berg corpus the topical dimension is obviously given : the alpine domain. Among the various text genres contained in the corpus we chose a set of 8 articles from one specific genre : mountaineering reports. We limited the geographical area of these reports to the Alps to be able to experiment with named entity recognition and geo-tagging (see (Volk *et al.*, 2010b)). The selected articles appeared in *Die Alpen* resp. *Les Alpes* in 1990 and 1991 (see table 2).

The selected articles in French and German do not contain foreign language sentences, except for one short Latin sentence ("Vivant amici montium !"). The interesting language mixture observed in the corpus is undesired in a parallel treebank. A German sentence in a French article would most likely result in alignments to "itself" in the parallel German text. In the unlikely case that it has a corresponding French sentence in the German article, it will still confuse the alignment system.

Year	Art. N°	French title	German title
1990	3	Une journée à Üschenen	Ein Tag in Üschenen
	6	Arête nord du Selbsanft	Erlebnis Selbsanft-Nordgrat
	13	Souvenirs du Piz Buin et du Piz Platta	Erinnerungen — Piz Buin und Piz Platta
	17	Deux fois le Rheinwaldhorn	Zweimal Rheinwaldhorn
	19	Wyss Wändli,chemin des souvenirs	Wyss Wändli — Weg der Erinnerungen
1991	2	Aux Piliers du Brouillard	In den Pfeilern des Brouillard
	4	Une odyssée alpine : la première traversée intégrale des 4000 suisses	Eine alpine Odyssee
	18	Un regard sur quelques aventures dans la région du Mont Blanc	Rückblick auf Abenteuer im Montblanc-Gebiet

TAB. 2 – Articles included in the Alpine Treebank

French Annotation The annotation of our treebanks is a semi-automatic process. We checked and corrected every step manually, from OCR output to morphosyntactic and functional annotation. Once the language of a sentence is identified, its token sequence is sent to the TreeTagger[5] configured for that particular language. We use the original English, German and Italian parameter files distributed with the TreeTagger. For French, we created a new parameter file by training the TreeTagger on an adapted version of the *Le Monde* corpus. To build the treebanks we loaded the tagged texts into Annotate[6], an interactive graphical tool that suggests constituent phrases and function labels based on a previously computed shallow parsing model (hidden Markov model).[7] We built the first French trees with the help of our adapted *Le Monde* model. After 200 sentences we generated a new parsing model based on our own set to avoid the noisy functional annotation of the larger set. We iteratively refined and completed the model by re-generating it periodically based on a larger set of annotated data.

Why and how did we adapt the French treebank (FTB) guidelines set by Abeillé's team for the annotation of the *Le Monde* corpus ? Our motivations were "to keep it simple" for the tokenization and "not to mix part-of-speech information with syntactic functions". Here we briefly discuss some issues in which the SMULTRON French treebanks differ from the FTB treebank. The main differences are the tokenization of the composed words (mots composés) –these having hyphens, apostrophes or simply spaces– , the simplification of the PoS tag and constituent schemes, eliminating the redundancy of some subcategorizations, and the explicit disambiguation of the function annotation.

At the morphosyntactic annotation level, we distinguish 13 lexical base categories further divided in subcategories resulting in 35 PoS tags in total. We did not want to introduce empty tokens, neither as trace nor as part of an agglutinated word. For example, the agglutinated preposition *au* is annotated as the simple preposition *à* with the PoS tag **P**, and the definite determiner *le* is not extracted. Elisions are explicited through secondary edges. Most expressions containing an apostrophe are split in separate tokens (*c'est* → *c' est*), a few exceptions registered in a stopword list are kept together (*aujourd'hui*, *prud'homme*). Unlike in FTB, the hyphenated words are not further decomposed and are considered as single words and annotated with a single PoS tag. There are two exceptions of this rule : clitics and prefixes. Postverbal clitics are split from the verb form ; but if the clitic itself contains an epenthetical *t*, then it is not further split (*a-t-il* → *a -t-il*). We simplified the PoS tags for clitics to separate the functional aspect from the morpho-syntactical annotation, reducing thus the three LeMonde tags to one : $CL_{\{sub|obj|ref\}} \rightarrow CL$. Frequent prefixes (*mi-*, *ex-*, *pseudo-*) are separated to mitigate the sparse data problem.

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[6] http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

[7] We are aware of the limitations of Annotate, but for our setting there is still no better alternative tool for constituent phrase annotation.

On the syntactic level, we group composed words in additional intermediate or pseudo-constituent categories : **CMP_N** (*sacs de couchage*), **CMP_P** (*à partir de*), **CMP_D** (*je ne sais quelle*), **CMP_ADV** (*en particulier*), etc. We also diverge from the general rule stated in the FTB guidelines and allow labelled edges between terminal and non-terminal nodes in one case : we annotate the functions under the verb group node **VN**. This way we eliminate the merged function labels like **A-OBJ/DE-OBJ** and keep the function information of each dependent element explicit.

We do not want to go into detail of the German treebank annotation as we have described this in (Samuelsson & Volk, 2006). Suffice it to say that we follow the German annotation guidelines of the TIGER treebank which allows crossing branches to represent variable constituent order and secondary edges for elliptical subjects. In a post-processing step we deepen the otherwise flat syntax trees (in deviation from TIGER) in order to create more complete and linguistically satisfying syntax trees.

French-German Alignment In order to align the sentences we applied the Bleualign algorithm[8] developed by Sennrich on our parallel corpus. (Sennrich & Volk, 2010) have shown that Bleualign applied on hard-to-align parallel texts outperformed other state-of-the-art methods. We discarded the many-to-many word alignments and kept the remaining word and sentence alignments. An annotator added the phrase alignments and completed resp. corrected the previously aligned sentences and words using TreeAligner[9], an open-source tool developed in-house to create, maintain and query parallel treebanks. Though the automatic alignment suggestions helped to speed up the manual process, the alignment was harder than expected due to the relative freedom of the translated texts. Such translation characteristics are a critical issue to consider while building a parallel treebank. Our experience shows that an average of 3-5 minutes is spent for annotating a sentence and the same amount of time for aligning a pair of corresponding trees.

	sentences	tokens	aligned	sentences	phrases	words	nodes (mixed)
FR	1075	22,085		1071	5327	11,801	
DE	1060	19,467		1055	4965	10,959	
FR-DE			pairs	1115	4995	12,016	17,488 (477)

TAB. 3 – Size of the Alpine treebank

Figure 1 contains an aligned tree pair. Node labels denote syntactic constituents like noun phrase (NP), adjective phrase (AP), and prepositional phrase (PP). Edge labels denote syntactic functions like subject (SUJ in French, SB in German), conjunct (CJ), or accusative object (OA in German). Green lines between the trees represent a close translation equivalences whereas red lines stand for approximate translation equivalences. The distinction between close and approximate equivalence is not always clear-cut, and we have compiled annotation guidelines with many examples to guide our annotators. We always take the perspective that aligned units should be reusable in a machine translation system outside of the given sentence context.

An issue to reconsider in future development is the French annotation of certain syntax structures, e.g. coordinations, that impede a maximal alignment with other language. It is impossible to align the first conjuncts, only the last element of the French coordination can be aligned to its German counterpart. The alignment of nominal phrases and also adjective phrases remains “incomplete”, the parallelism potential is not fully exploited.

Parallel treebanks allow precise queries over aligned syntax structures. For example, it is possible to query for French relative clauses which have aligned German noun phrases. Our query tool supports the full range of treebank queries combined with alignment constraints. Treebank queries comprise dominance relations (e.g. category X dominates part-of-speech Y), precedence relations (e.g. word X precedes part-of-speech Y), tree predicates (e.g. category X has 5 daughters) and any combinations thereof. Such queries can be formulated over both trees in a bilingual tree pair and may then be refined with alignment constraints. This allows searches for and insights into translation correspondences that were hitherto impossible.

4 Conclusion

We have presented our multilingual and partly parallel corpus of alpine texts. This corpus contains a large number of articles in French and German and is thus a welcome addition to parallel corpora for different applications in language technology. We currently use it to investigate how to best combine corpora for domain-specific SMT.

[8] http://github.com/rsennrich/bleualign
[9] http://www.cl.uzh.ch/kitt/treealigner

Text+Berg is an ongoing project in which we constantly refine the corpus by removing OCR errors and adding more layers of annotation. We also plan to increase the corpus size. To this end we are currently processing the *Echo des Alpes*, the yearbooks 1872-1923 of the French section of the Swiss Alpine Club. Not only will the corpus grow but this additional subcorpus will also balance the German majority with a new French counterpart : 6.5 million tokens from 5900 articles. And neither last nor least, we will give stronger support to English in our corpus by including texts from the British Alpine Club. As a next step we plan to automatically build a large parallel treebank on the parallel part of the Text+Berg corpus following (Sennrich & Volk, 2010), (Zhechev, 2009) and (Tiedemann & Kotzé, 2009). On the other hand, we continue to manually annotate and align new treebanks covering other domains and genres.

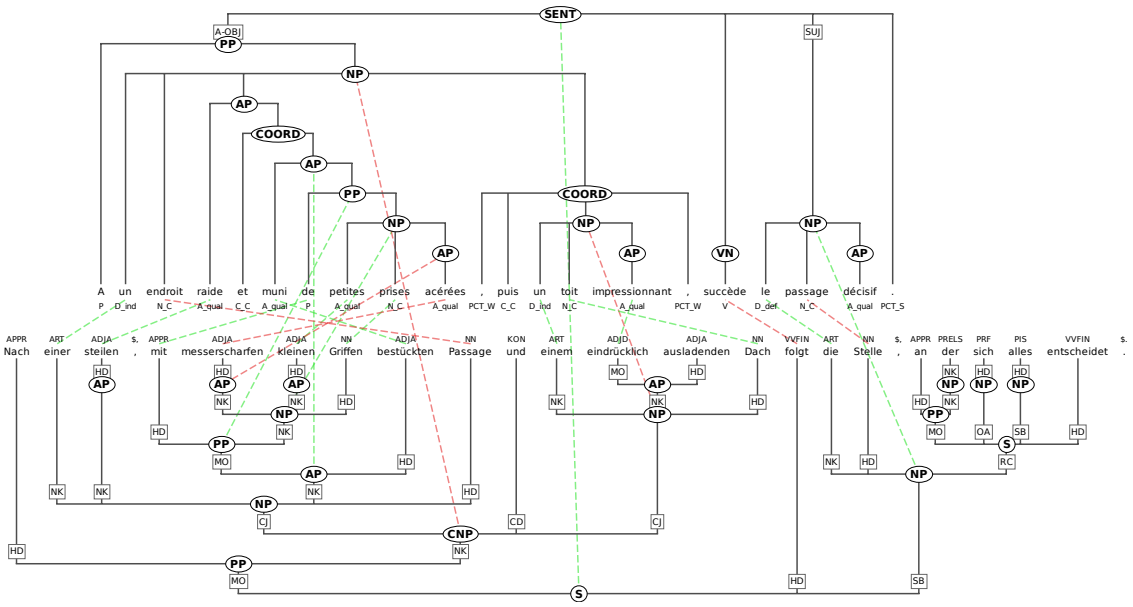

FIG. 1 – Aligned parallel French-German trees

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Treebanks. Building and Using Parsed Corpora*, chapter 13 : Building a treebank for French, p. 165–188. Text, Speech and Language Technology. Kluwer Academic Publishers : Dordrecht.

SAMUELSSON Y. & VOLK M. (2006). Phrase alignment in parallel treebanks. In *5th Workshop on Treebanks and Linguistic Theories*.

SENNRICH R. & VOLK M. (2010). MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of AMTA*, Denver.

TIEDEMANN J. & KOTZÉ G. (2009). Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories*, p. 197–208, Milano.

VOLK M., BUBENHOFER N., ALTHAUS A., BANGERTER M., FURRER L. & RUEF B. (2010a). Challenges in building a multilingual alpine heritage corpus. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

VOLK M., GÖHRING A. & MAREK T. (2010b). Combining parallel treebanks and geo-tagging. In *Fourth Linguistic Annotation Workshop (LAW IV)*.

VOLK M., GÖHRING A., MAREK T. & SAMUELSSON Y. (2010c). SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank. http ://www.cl.uzh.ch/research/paralleltreebanks_en.html. An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments.

ZHECHEV V. (2009). *Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System*. PhD thesis, School of Computing at Dublin City University.