Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC

Stephen Grimes, Xuansong Li, Ann Bies, Seth Kulick, Xiaoyi Ma, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA USA
E-mail: { sgrimes, xuansong, bies, skulick, xma, strassel}@ldc.upenn.edu

Abstract

This contribution describes an Arabic-English parallel word aligned treebank corpus from the Linguistic Data Consortium that is currently under production. Herein we primarily focus on efforts required to assemble the package and instructions for using it. It was crucial that word alignment be performed on tokens produced during treebanking to ensure cohesion and greater utility of the corpus. Word alignment guidelines were enriched to allow for alignment of treebank tokens; in some cases more detailed word alignments are now possible. We also discuss future annotation enhancements for Arabic-English word alignment.

1. Introduction

1.1 Parallel Treebanks

Multiple annotation of corpora is common in the development of computational linguistic language resources. Additional annotation increases potential information extraction from a given resource. For example, many existing parallel corpora have been developed into parallel treebanks, and for several language pairs there exist parallel treebank corpora. Parallel treebank corpora are parallel texts for which there exist manual parses for both languages (and possibly POS tags also). Examples include Czech-English (Hajic et al., 2001), English-German (Cyrus et al., 2003), English-Swedish (Ahrenburg, 2007), Swedish-Turkish (Megyesi et al., 2008), Arabic-English (Maamouri et al., 2005; Bies, 2006), Chinese-English (Palmer et al., 2005; Bies et al., 2007). The latter corpora produced by LDC are of particular note due to their high data volume.

Parallel word-aligned treebank corpora appear to be rare, and their scarcity is likely due to their being very resource-intensive to create. The most prominent related corpus is called SMULTRON and is a parallel aligned treebank corpus for one-thousand English, Swedish, and German sentences (Gustafson-Capkova et al., 2007). In SMULTRON, alignment is pairwise between each of the component languages, and annotation permitted between syntactic categories and not exclusively between words.

1.2 Current Project

The present paper discusses key points in creating an Arabic-English parallel word-aligned treebank corpus. We have also included a brief description of this corpus in the LREC 2010 Language Resource Map.

As shown in Table 1, releases for this corpus began in 2009, and to date more than 325,000 words of Arabic and the corresponding English translation have been treebanked and word aligned. Each release includes data from one or more genre: newswire (NW), broadcast news transcripts (BN), or online web resources such as blogs (WB).

1.3 Organization of the Paper

The paper is structured as follows. Section 2 discusses development of Arabic and English treebanks. Section 3 discusses word alignment at LDC. Section 4 addresses issues faced in combining treebank and word alignment annotation. Section 5 has information about the corpus structure and how to use the data. Section 6 provides a critical analysis and discussion of future directions.

Release date	Genre	Words	Tokens	Sentences
4/9/2009	NW	9191	13145	382
9/21/2009	NW	182351	267520	7711
9/21/2009	BN	89213	115826	4824
10/24/2009	NW	16207	22544	611
10/24/2009	WB	6656	9478	288
1/29/2010	BN	9930	12629	705
1/29/2010	WB	12640	18660	565
Total		326188	459802	15086

Table 1. Annotation volume as of May 2010. Figures reported for words and tokens refer to the Arabic source.

2. Development of Parallel Treebanks

The path towards construction of the resource under discussion could be considered to begin with the Arabic Treebank (ATB) corpus (Maamouri et al., 2005). Translation of the Arabic to English created parallel texts, and when the English-Arabic Translation Treebank (EATB) (Bies, 2006) is used in conjunction with the ATB, this serves as an English-Arabic parallel treebank corpus. Please refer to documentation released with these corpora for additional discussion concerning construction, annotation guidelines, and quality control efforts that went into creating the individual treebanks.

In developing parallel treebanks, care must be taken to ensure sentence segments remain parallel from the original parallel corpus. Arabic sentences are often translated as multiple English sentences. Hence one Arabic tree may correspond to multiple English trees, and occasionally effort is required to enforce that sentence segments remain parallel. For a similar project involving an English-Chinese parallel word-aligned treebanked corpus, English and Chinese treebanking were performed independently at different locations, and the resulting corpora were only weakly parallel; an automatic sentence aligner was required to re-establish the parallel texts. We used Champollion, a lexicon-based sentence aligner for robust alignment of the noisy data (Ma, 2006). Such a tool may be necessary for others creating parallel aligned treebank corpora if the data inputs are not already sentence-wise parallel.

The Arabic Treebank (ATB) distinguishes between source and treebank tokens. While source tokens are generally whitespace-delimited words, the treebank tokens are produced using a combination of SAMA (Maamouri et al., 2009) for morphological analysis, selection from amongst alternative morphological analyses, and finally splitting of the source token into one or more treebank tokens based on clitic or pronoun boundaries.

For release as part of this corpus, the ATB and EATB are provided in Penn Treebank format (Bies et al., 1995). The trees are unmodified from ATB/EATB releases except that the tokens were replaced with token IDs. This structure is discussed in greater detail in Section 5.

3. Word Alignment Annotation

At the LDC, word alignment is a manual annotation process that creates a mapping between words or tokens in parallel texts. While automatic or semi-automatic methods exist for producing alignments, we avoid these methods. Manual alignment serves as a gold standard for training automatic word alignment algorithms and for use in machine translation (c.f. Melamed 2001, Véronis and Langlais 2000), and it is desirable that annotator decisions during manual alignment not be biased through use of partially pre-aligned tokens. It is felt that annotators may accept the automatic alignment and also lower annotator agreement at the same time.

Using higher-quality manual alignment data for training data results in better machine translations. Fossum, Knight, and Abney (2008) showed that using Arabic and English parsers or statistical word alignment tools such as GIZA++ instead of gold standard annotations contributes to degradations in training data quality that significantly impact BLEU scores for machine translation. While automatic parsing and word aligning have their place in NLP toolkits, use of manually-annotated training data is always preferred if annotator resources are available.

3.1 Word Alignment Annotation Guidelines

LDC's word alignment guidelines are adapted from previous task specifications including those used in the BLINKER project (Melamed 1998a, 1998b). Single or multiple tokens (words, punctuation, clitics, etc.) may be aligned to a token in the opposite language, or a given token may be marked as not translated. Early LDC Arabic-English word alignment releases as part of the DARPA GALE program were generally based on whitespace tokenization.

Word alignment guidelines serve to increase annotator agreement, but different word alignment projects may have unique guidelines according to what is deemed translation equivalence. For example, are pronouns permitted to be aligned to proper nouns with which they are coindexed? Our point here is to encourage the corpus user to explore alignment guidelines in detail to better understand the task.

3.3 Word Alignment and Tagging Tool

Word alignment is performed on unvocalized tokens rendered in Arabic script. LDC's word alignment tool allows annotators to simultaneously align tokens and tag them with meta data or semantic labels. A screenshot of the tool is shown in Figure 1.

The navigation panel on the right side of the software displays original (untokenized) source text to help annotators understand the context of surrounding sentences (which aids in, for example, anaphora resolution). Having untokenized source text also aids in resolving interpretation ambiguities that would arise if annotators could only see tokenized, unvocalized script.

3.3 Additional Tagging for Word Alignment

In addition to part-of-speech tags produced as part of treebank annotation, word alignment annotators have the option of adding certain language-specific tags to aid in disambiguation. A tagging task for Arabic-English has recently been added to the duties of word alignment annotators, and it is described as follows.

For unaligned words or phrases having locally-related constituents to which to attach, they are tagged as "GLU" (i.e., "glue"). This indicates local word relations among dependency constituents. The following are some cases in which the GLU tag would be used:

-English subject pronouns omitted in Arabic.

-Unmatched verb "to be" for Arabic equational sentences.

-Unmatched pronouns and relative nouns when linked to their referents.

-Unmatched possessives ('s and ') when linked to their possessor.

-When a preposition in one language has no counterpart, the extra preposition attached to the object is marked GLU.

-Two or more prepositions in one language while there is one preposition in the other side; the unmatched preposition would be tagged as GLU.

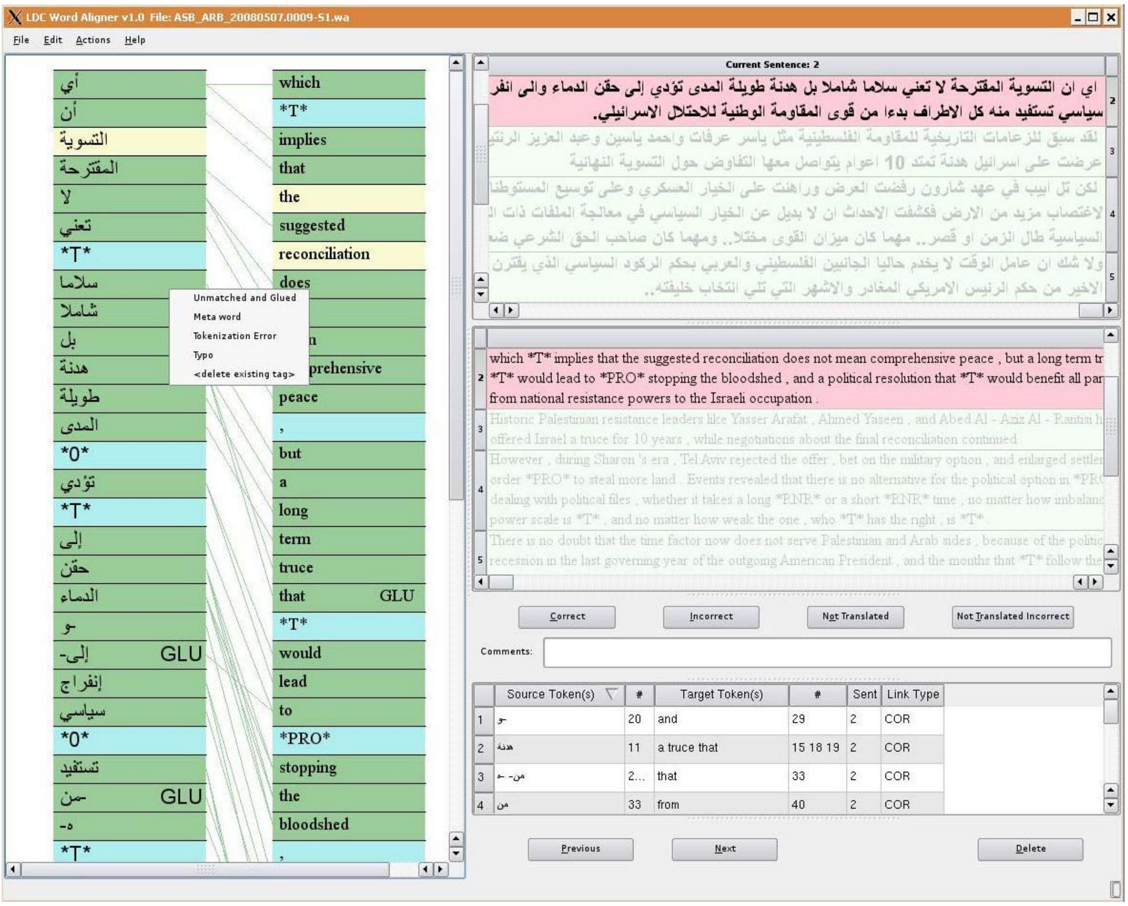
It is hoped that the presence of the GLU tag provides a clue in understanding morphology better, and we will continue to explore using additional tags for this task.

4. Uniting Treebank and Word Alignment Annotation

This section describes efforts to join treebank and word alignment annotation.

4.1 Order of Annotation

The order of annotation in creating a parallel word-aligned treebank corpus is important. From the parallel corpus, the sentences can first be treebanked or word aligned. If word alignment was to proceed first, the tokens used for word alignment would serve as input to treebanking. However, treebank tokenization includes morphosyntactic analysis, and hence treebank tokenization is only determined manually during treebank annotation. For this reason, the preferred workflow is to only perform word-alignment annotation after experienced treebank annotators have fixed tokenization,



2 اي ان التسوية المقترحة لا تعني سلاما شاملا بل هدنة طويلة المدى تؤدي إلى حقن الدماء والى انفر
سياسي تستفيد منه كل الاطراف بدءا من قوى المقاومة الوطنية للاحتلال الاسرائيلي.

3 لقد سبق للزعامات التاريخية للمقاومة الفلسطينية مثل ياسر عرفات واحمد ياسين وعبد العزيز الرنتي
عرضت على اسرائيل هدنة تمتد 10 اعوام يتواصل معها التفاوض حول التسوية النهائية

4 لكن تل ابيب في عهد شارون رفضت العرض وراهنت على الخيار العسكري وعلى توسيع المستوطنا
لاغتصاب مزيد من الارض فكشفت الاحداث ان لا بديل عن الخيار السياسي في معالجة الملفات ذات ال
السياسية طال الزمن او قصر.. مهما كان ميزان القوى مختلا.. ومهما كان صاحب الحق الشرعي ضع

5 ولا شك ان عامل الوقت لا يخدم حاليا الجانبين الفلسطيني والعربي بحكم الركود السياسي الذي يقترن
الاخير من حكم الرئيس الامريكي المغادر والاشهر التي تلي انتخاب خليفته..

2 which *T* implies that the suggested reconciliation does not mean comprehensive peace , but a long term tr
T would lead to *PRO* stopping the bloodshed , and a political resolution that *T* would benefit all par
from national resistance powers to the Israeli occupation .

3 Historic Palestinian resistance leaders like Yasser Arafat , Ahmed Yaseen , and Abed Al - Aziz Al - Rantisi h
offered Israel a truce for 10 years , while negotiations about the final reconciliation continued .

4 However , during Sharon 's era , Tel Aviv rejected the offer , bet on the military option , and enlarged settler
order *PRO* to steal more land . Events revealed that there is no alternative for the political option in *PR
dealing with political files , whether it takes a long *RNR* or a short *RNR* time , no matter how imbalanc
power scale is *T* , and no matter how weak the one , who *T* has the right , is *T* .

5 There is no doubt that the time factor now does not serve Palestinian and Arab sides , because of the politic
recession in the last governing year of the outgoing American President , and the months that *T* follow the

	Source Token(s)	#	Target Token(s)	#	Sent	Link Type
1	-و	20	and	29	2	COR
2	هدنة	11	a truce that	15 18 19	2	COR
3	من- -ه	2...	that	33	2	COR
4	من	33	from	40	2	COR

Figure 1. The PyQt-based tool used at LDC for word alignment annotation and tagging.

and it is this development trajectory we assume for the remainder of the paper.

4.2 Tokenization Modification

The word alignment guidelines were adapted so that annotation would be based on the treebank tokens instead of on source, whitespace tokens. As illustrated by the following examples, finer alignment distinctions may be made when pronouns are considered independent tokens. The example below appears in the Buckwalter transliteration[1] for convenience, but please note that the bilingual annotators work only with Arabic script.

Source:	فزجّوه بالسجن
Transliterated:	fa+ zaj~uwA +h b+ Alsijn
Morpheme gloss:	and sent.3P him to jail
Gloss:	"They sent him to jail."

Source:	سيّارته معطّلة
Transliterated:	say~Arap+h muEaT~alap
Morpheme gloss:	car his broken
Gloss:	"His car is broken."

In each case the "h" morpheme corresponding to third person singular is now considered an independent token and can be aligned to English "him" or "his" in the examples. Under previous Arabic-English word alignment guidelines , English "him" and "his" would have been aligned with the Arabic verb.

4.3 Empty Categories

In transitioning to word alignment on treebank tokens, all leaves of the syntax tree — including all Empty Categories — are considered to be tokens. This interpretation as tokens differs slightly from ATB- and EATB-defined treebank tokens which do not include the Empty Category markers such as traces, empty complementizers, and null pro markers.

Our word alignment guidelines currently dictate that all Empty Category tokens are annotated as "not translated." One could imagine amending guidelines to allow for the alignment of Empty Category markers to pronouns in the translation. This is not currently being practiced. The primary reason for including Empty Categories as tokens for word alignment is to ensure that, for each language, the number of tree leaves is identical to the number of word alignment tokens. This requirement simplifies somewhat the data validation process.

4.4 Data Validation

Validation of the data structures have both manual and automatic components.

4.4.1 Treebank validation

Throughout the Treebank pipelines, there are numerous stages and methods of sanity checks and content validation, to assure that annotations are coherent, correctly formatted, and consistent within and across annotation files, and to confirm that the resulting annotated text remains fully concordant with the original transcripts (for Arabic) or translations (for English), so that cross-referential integrity with the original data and with English translations is maintained.

For both Arabic Treebank and English Treebank, quality control passes are performed to check for and correct errors of annotation in the trees. The Corpus Search tool[2] is used with a set of error-search queries created at LDC to locate and index a range of known likely annotation errors involving improper patterns of tree structures, node labels, and the correspondence between part-of-speech tags and tree structure. The errors found in this way are corrected manually in the treebank annotation files.

In addition, the Arabic Treebank (ATB) closely integrates the Standard Arabic Morphological Analyzer (SAMA) into both the annotation procedure and the integrity checking procedure. The interaction between SAMA and the Treebank is evaluated throughout the workflow, so that the link between the Treebank and SAMA is as consistent as possible and explicitly notated for each token.

For details on the integration between the ATB and SAMA, along with information about the various forms of the tokens that are provided, see Kulick, Bies and Maamouri (2010). For a general overview of the ATB pipeline, see Maamouri, et al. (2010).

4.4.2 Word alignment validation

For word alignment, it is verified that all delivery files are well-formed. It is ensured that all tokens receive some type of word alignment annotation.

4.4.3 Validation of parallel word-aligned treebanks

To ensure consistency of the parallel aligned treebank, we verify that the set of tokens referenced by the treebank files coincides with the same set of tokens appearing the token and word alignment files.

5. Using the Corpus

This section provides information about the file format of the word-aligned treebanked data we are releasing. A typical release will contain seven files for each source document

-- Arabic source, collected from newswire, television broadcast, or on the web
-- English translation of Arabic source
-- Tokenized Arabic, resulting from treebank annotation
-- Tokenized English, resulting from treebank annotation
-- Treebanked Arabic
-- Treebanked English
-- Word alignment file

The parallel treebank is a standoff annotation with multiple layers of annotations with upper layer annotation referring to lower layer data (using character offsets). The diagram in Figure 2 shows the dependencies between files in the release.

[1] We use the Buckwalter transliteration. Details are available at http://www.qamus.org/transliteration.htm.

[2] CorpusSearch is freely available at http://corpussearch.sourceforge.net

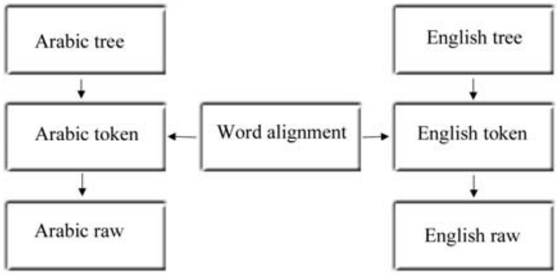


Figure 2. File structure illustration

The word alignment file and the Arabic and English tree files have token numbers which reference the Arabic and English token files. Within the token files, each token number for each sentence is expanded to give additional information. For each token in the English token files, the token number is listed, followed by a character range in the raw file to which the token corresponds, and then finally the token itself. For Arabic, multiple versions of each token are provided (unvocalized, vocalized, input string) and in multiple formats (Arabic script, Buckwalter transliteration).

We considered distributing the corpus in a single XML-based file. We felt the present structure has the following advantages:
-- the format of each type of file (raw, tokenized, tree, wa) is not modified and hence the same tools researchers wrote before can still be used;
-- the data are more easily manipulated; with XML it is necessary to fully parse the xml files for even trivial tasks;
-- it is easier and less error-prone to put the package together using separate files then using xml; and
-- separate files are more human readable.

6. Discussion

Annotator agreement for the Arabic-English word alignment task is approximately 85% after first pass annotation and higher after a quality round of annotation. In the future we plan to add additional morphosyntactic or semantic tags to the word alignment portion of the task.

We are also investigating methods for improving automatic and semi-supervised error detection. We wish to flag statistically unlikely alignments for human annotator review. Additionally, through incorporating phrase structure from treebank annotation, we might examine alignments which cross certain phrase boundaries.

7. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

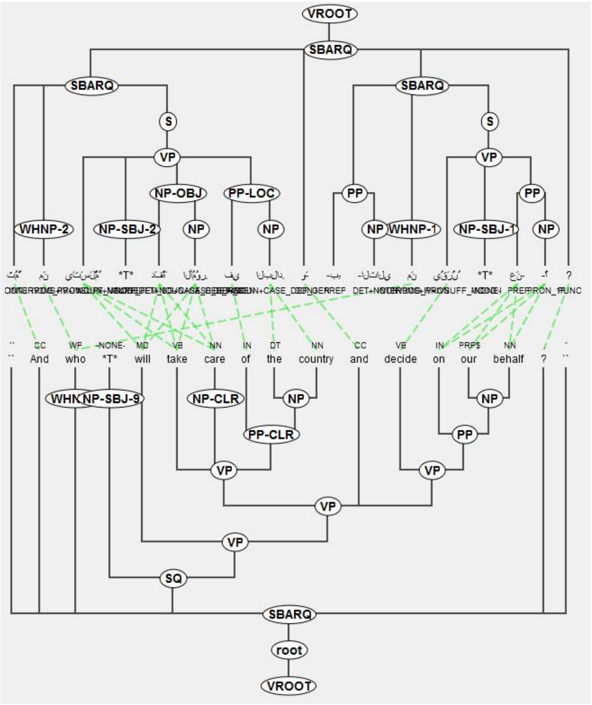


Figure 3. A view of Arabic (above) and English (below) word-aligned treebanks as displayed by TreeAligner[3].

8. References

Bies, A., Ferguson, M., Katz, K, and MacIntyre, R. (1995). Bracketing guidelines for treebank II style, Penn Treebank Project. University of Pennsylvania technical report.

Bies, A. (2006). English-Arabic Treebank v 1.0. LDC Cat. No.: LDC2006T10.

Bies, A., Palmer, M., Mott, J., and Warner, C. (2007). English Chinese Translation Treebank v 1.0. LDC Cat. No.: LDC2007T02.

Cyrus, L., Feddes, H., and Schumacher, F. (2003). Fuse – a multilayered parallel treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*.

Fossum, V., Knight, K., and Abney S. (2008). Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation. In *Proceedings of Third Workshop on Statistical Machine Translation*, p.44-52, Columbus, June 2008. Assocation for Computational Linguistics.

Gustafson-Capkova, S., Samuelsson, Y., and Volk, M. (2007). SMULTRON (version 1.0) - The Stockholm MULtilingual parallel TReebank. Department of Linguistics, Stockholm University, Sweden.

Hajic, J., Hajicova, E., Pajas, P. , Panevova, J., Sgall, P. and Vidova Hladka, B. (2001). The Prague Dependency Treebank 1.0 CDROM. LDC Cat. No. LDC2001T10.

Kulick, S., Bies, A., and Maamouri, M. (2010). Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank. In *Proceedings of the Seventh*

[3] http://kitt.cl.uzh.ch/kitt/treealigner

International Conference on Language Resources and Evaluation (LREC 2010).

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation.*

Maamouri, M., Bies, A., Buckwalter, T., and Jin, H. (2005). Arabic Treebank: Part 1 v 3.0 (POS with full vocalization + syntactic analysis). LDC Cat. No.: LDC2005T02.

Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D., and Ciul, M.. (2010). From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010).*

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., and Kulick, S. (2009). *LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.0*. LDC Catalog No.: LDC2009E44. Special GALE release to be followed by a full LDC publication.

Megyesi, B., Dahlqvist, B., Pettersson, E., and Nivre, J. (2008). Swedish-Turkish Parallel Treebank. In *Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2008).*

Melamed, D.I. (2001). Empirical Methods for Exploiting Parallel Texts. MIT Press.

Melamed, D.I. (1998a). Annotation Style Guide for the Blinker Project. University of Pennsylvania (IRCS Technical Report #98-06).

Melamed, D.I. (1998b). Manual Annotation of Translational Equivalence: The Blinker Project. University of Pennsylvania (IRCS Technical Report #98-07).

Palmer, M., Chiou, F.-D., Xue, N., and Lee, T.-K. (2005) LDC2005T01, Chinese Treebank 5.0.

Véronis, J. and Langlais, P. (2000). Evaluation of Parallel Text Alignment Systems -- The ARCADE Project. In J. Véronis (ed.) *Parallel Text Processing, Text, Speech and Language Technology*. ordrecht, The Netherlands: Kluwer Academic Publishers, pp. 369-388.