An Arabic Dependency Treebank in the Travel Domain

Dima Taji, Jamila El Gizuli, Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
{dima.taji,nizar.habash}@nyu.edu

Abstract

In this paper we present a dependency treebank of travel domain sentences in Modern Standard Arabic. The text comes from a translation of the English equivalent sentences in the Basic Traveling Expressions Corpus. The treebank dependency representation is in the style of the Columbia Arabic Treebank. The paper motivates the effort and discusses the construction process and guidelines. We also present parsing results and discuss the effect of domain and genre difference on parsing.

Keywords: Arabic, Dependency, Treebank, Travel, Tourism

1. Introduction

Treebanks, or annotated corpora, are essential for Natural Language Process (NLP) tasks. Such tasks include building lexicons, inferencing grammars, and creating computational analyzers, which can all be improved by creating treebanks with different kinds of linguistic annotations (Abeillé, 2012). Treebanks with rich and good quality annotation are very expensive resources to create. They require a large number of man-hours to create and audit.

Treebanks can be in multiple genres, or genre-specific.[1] However, there is a tradeoff between the cost of the size, the diversity of a corpus, and having enough content in one genre or domain to be able to make generalizations. As a result, many treebanks tend to be predominantly of one specific genre, but may add some samples of other genres. For example, the Hindi/Urdu Treebank (Bhat et al., 2017) is predominantly in the news domain with 85.3% of its sentences coming from news articles, and only 14.7% from other domains (9.7% from conversations, and 5% from the travel domain). Webber (2009) shows that the Penn Treebank (Marcus et al., 1994) consists of 90.1% news articles, 4.9% essays, 2.6% summaries, and 2.4% letters, and it is still considered to be a news domain treebank. Similarly, Maamouri et al. (2010) demonstrate that the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) consists of 39.9% newswire text, 28.2% broadcast news, 18.6% broadcast conversation in both Standard and Dialectal Arabic, and 13.3% web texts.

In this paper we describe a small Modern Standard Arabic (MSA) treebank, created using a travel corpus. This treebank will be the seed of a larger multi-genre, and multi-dialect Arabic treebank. The corpus we are using is part of an MSA translation by Eck and Hori (2005) of the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2007), henceforth MSABTEC. As far as we know, there is no treebank based on this corpus.

[1]The terms *domain*, *genre*, *topic* and *style* have been discussed a lot in the field (Lee, 2002; Van der Wees et al., 2015; Ide and Pustejovsky, 2017), and many authors discussed their ambiguous and overlapping use. For the rest of this paper we use the term travel domain, following Takezawa et al. (2007) whose corpus was the basis for the translated corpus we treebank.

In Section 2, we discuss related work followed by a description of the corpus we annotate in Section 3. In Section 4, we discuss the annotation format; and in Section 5 the annotation process. Finally, we present some results on benchmarking parsing on our corpus and a comparison with a major news-domain Arabic treebank in Section 6.

2. Related Work

BTEC is a collection of conversational phrases that cover various situations in the travel domain in Japanese, and their translations into English and Chinese (Takezawa et al., 2007). The sentences in the corpus were collected from bilingual travel experts, and were based on their experience rather than being transcribed. The corpus was later translated into more languages including Arabic (Eck and Hori, 2005), where it was used for evaluating machine translation systems.

Another treebank that included phrases from the travel domain is the Hindi/Urdu treebank (Bhat et al., 2017). Even though the majority of the treebank comes from news sources, it contains 15K words, making up 1,058 sentences relating to heritage and tourism. This part of the data was specifically added to counteract the bias that could result from using data in one specific domain, news in this instance. The treebank contains dependency, phrase-structure, and PropBank-inspired (Kingsbury et al., 2002) annotations.

The Penn Treebank is a well known resource, that contains phrases mostly from the news domain. The treebank was annotated for genres as part of the Penn Discourse Treebank (Miltsakaki et al., 2004), and Webber (2009) shows that the different genres can have different characteristics.

The Penn Arabic Treebank (PATB) is the primary treebank for work on Arabic syntactic analysis. It uses a phrase-structure representation, but has been converted to other dependency formalisms (Habash and Roth, 2009; Taji et al., 2017). The PATB contains various parts that come from different domains and resources. PATB comes in 12 parts (Diab et al., 2013), that are mostly from news or web sources (Maamouri et al., 2010). Other related treebanks were developed by the Linguistic Data Consortium

(LDC) in various dialects such as Egyptian (Maamouri et al., 2012), and Levantine (Maamouri et al., 2006), where the data came from transcribing recorded conversations.

The first dependency Arabic treebank was the Prague Arabic Dependency Treebank (PADT) (Hajič et al., 2004). It employed a multi-level description scheme for functional morphology, analytical dependency syntax, and tectogrammatical representation of linguistic meaning.

Another large Arabic treebank is the Columbia Arabic Treebank (CATiB) (Habash and Roth, 2009). CATiB has around 250K words that were annotated directly in its dependency representation, in addition to a fully converted version of the PATB (PATB-CATiB). CATiB focuses on news domain text in Standard Arabic. Most recently, Taji et al. (2017) converted the PATB into the formalism of the Universal Dependency (UD) project (Nivre et al., 2016) via an intermediate step of mapping to CATiB dependencies.

The Quran Corpus is another important Arabic syntactic corpus of the very specific genre of holy scripture (Dukes and Buckwalter, 2010). It has its own representation scheme which is a hybrid dependency and constituency.

In this work, we annotate in the format of the CATiB treebank and compare to UD representations. And we present a comparison with the news domain as captured in the PATB.

3. Our Corpus

For our corpus, we selected the MSA translation of BTEC (Eck and Hori, 2005). Our selection contains 2,000 sentences making a total of 15,929 words (7.9 words/sentence). The sentences chosen are the same as those in CORPUS-25 from the Multi Arabic Dialect Applications and Resources (MADAR) project (Bouamor et al., 2018). The text of the corpus, coming from BTEC, is full of travel related expressions such as inquiring about the prices of hotel rooms, asking for directions, requesting help, ordering food, etc. Being conversational, it also has a high percentage of first and second person pronouns and conjugations. Below are examples of sentences from MSABTEC:

- أحتاج إلى طبيب. *ÂHtAj ĂlY Tbyb.*[2] 'I need a doctor.'
- كريمة وسكر؟ *krymħ wskr?* 'Cream and sugar?'
- أين أقرب محل جزارة؟ *Âyn Âqrb mHl jzArħ?* 'Where is the nearest butcher?'

4. Annotation Format

To maximize compatibility with previous efforts, we followed the Columbia Arabic Treebank (CATiB) (Habash and Roth, 2009) annotation guidelines, and tokenization schemes used by previous Arabic treebanks. We chose this format because it uses traditional Arabic grammar as the inspiration for its relational labels and dependency structure (Habash and Roth, 2009), making it intuitive for Arabic speakers, and allowing for faster annotation. In addition, this format can be automatically enriched with more morphological features (Alkuhlani et al., 2013), and converted into other dependency formats such as the Universal Dependency format (Taji et al., 2017). Except for a number of minor specifications for some new syntactic constructions, there was no change to the guidelines for tokenization, part-of-speech (POS) tag set, and relations.

4.1. Tokenization

The tokenization followed in the treebank creation is the same tokenization scheme used in PATB. This scheme tokenizes all the clitics, except for the definite article +ال *Al+* 'the' (Habash, 2010). The 2,000 sentences in our corpus consist of 18,628 tokens (manually checked).

4.2. Annotation Scheme

For our treebank, we followed the CATiB dependency annotation scheme. This scheme is designed to be speedy for annotation, and intuitive for Arabic speakers. We also used the guidelines that were prepared for the CATiB annotation project (Habash et al., 2009).

4.2.1. POS Tags

The CATiB annotation scheme uses six POS tags which are **NOM** for all nominals excluding proper nouns, **PROP** for proper nouns, **VRB** for active-voice verbs, **VRB-PASS** for passive-voice verbs, **PRT** for particles, which include prepositions and conjunctions, and **PNX** for punctuation marks.

4.2.2. Relations

There are eight relations used in the CATiB scheme: **SBJ** for the subjects of verbs and the topics of simple nominal sentences; **OBJ** for the objects of verbs, prepositions, or deverbal nouns; **TPC** for the topics of complex nominal sentences which contain explicit pronominal referents; **PRD** for the complements of the extended copular constructions; **IDF** for marking the possessive nominal construction (*idafa*); **TMZ** for marking the *specification* nominal construction (*tamyiz*); **MOD** for general modification of verbs or nominals; and, finally, **—** for marking flat constructions such as first-last proper name sequences.

4.2.3. Syntactic Structures

Since the original CATiB treebank, as with the Penn Arabic treebank, was focused on the news genre, there were many syntactic constructions that MSABTEC introduced that needed special attention. In particular, there was an abundance of interrogatives, and first and second person statements in MSABTEC compared to CATiB. To address these constructions, additional guideline specifics and clarifications were added. All of these extensions followed naturally from the spirit of the original guidelines. For example, an interrogative pronoun such as من *man* 'who/whom' is often sentence-initial, but it can be the subject or the object of a verb: من سمع +ك؟ *man samiςa +ka?* 'who heard you?' versus من سمعت؟ *man samiςta?* 'whom did you hear?'. Similarly, in Figure 1 (C), the interrogative adverb أين *Âyn* 'where' is treated as the predicate head of a copular sentence since that is the syntactic role of the answer to the question. For another common example in this genre, single word interjections such as آسف *Āsf* 'sorry'

[2] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

or شكراً *škrAã* 'thanks' are treated as independent sentence trees that attached directly to the main root of the sentence they appear in.

4.3. Interface

The annotation was done using the TrEd annotation interface (Pajas, 2008), which was also used by Habash and Roth (2009) for CATiB annotation.

Figure 1 illustrates the annotation scheme of three examples from the MSABTEC Treebank in the CATiB format in which they were annotated. We also provide, for comparison, the analysis in the increasingly popular Universal Dependency representation (Nivre et al., 2016; Taji et al., 2017).

5. Annotation Process

The annotation process we followed in the preparation of this treebank is the same process described by Habash and Roth (2009), which consisted of the following steps: *(a) Automatic Tokenization and POS Tagging*, *(b) Manual Tokenization Correction*, *(c) Automatic Parsing*, and *(d) Manual Annotation*. In this section, we discuss what we did for these steps as well as report on annotator(s), speed, and inter-annotator agreement.

5.1. Annotator(s)

Due to the relatively small size of our treebank, we had only one annotator working on the task. Our annotator is an educated native Arabic speaker, who was trained on the CATiB scheme and the use of TrEd as part of her work on the original CATiB project (Habash and Roth, 2009). To evaluate inter-annotator agreement, we worked with a second annotator who was asked to annotate a small part of the treebank (see below).

5.2. Automatic Tokenization and POS Tagging

We used MADAMIRA (Pasha et al., 2014) to tokenize and POS tag the input sentences. We used MADAMIRA's configuration for PATB tokenization and CATiB POS tags.

5.3. Manual Tokenization Correction

Our annotator then manually checked and fixed all of the tokenization errors. This also included the correction of typos and spelling changes resulting from wrong automatic analysis. Overall there were 2.8% tokenization errors, which is higher than MADAMIRA's reported tokenization error rate (around 1.1%). The increase is most likely due to the difference in genre between the data used to train MADAMIRA and our corpus.

5.4. Automatic Parsing

We ran the data with the fixed tokenization through the CamelParser (Shahrour et al., 2016), which is trained on the gold CATiB representation of the training data from the PATB parts 1, 2, and 3, according to the splits proposed by Diab et al. (2013). We present automatic parsing quality results in Section 6.2.

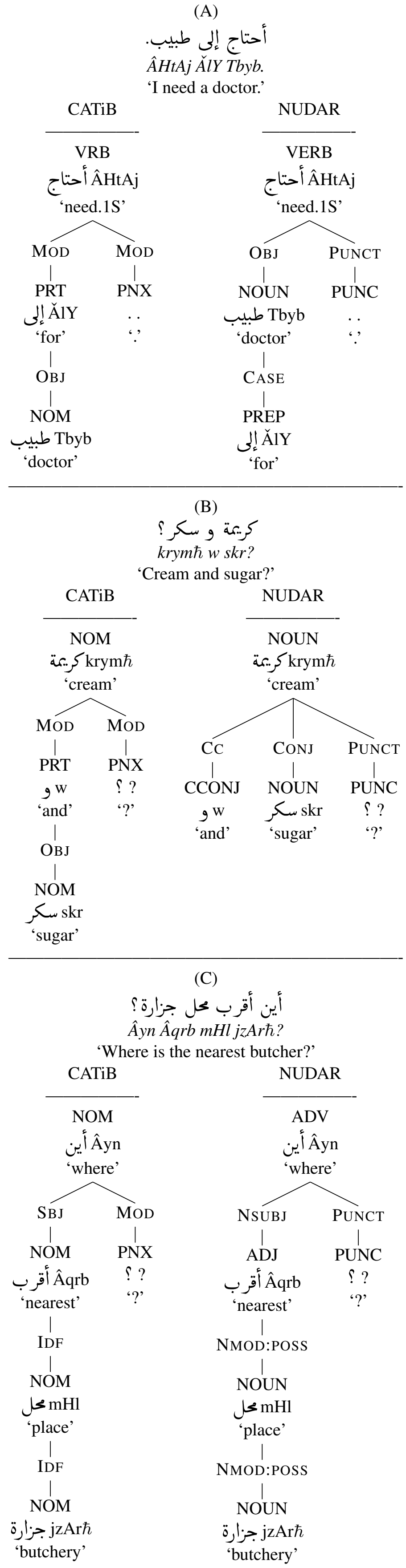

Figure 1: The structures for example trees from the MSABTEC Treebank in CATiB format, and their counterpart in the Arabic Universal Dependency (NUDAR) format (Taji et al., 2017).

5.5. Manual Annotations

The output of the automatic parsing was given in TrEd's *.fs* format to the annotator to manually fix the POS tags, the relation labels, and the syntactic structures of the trees.

5.6. Annotation Speed

The manual fixing of the tokenization took the annotator 10 hours of work at the speed of 1,593 words/hour. The manual correction of the parsed trees (POS, relations, and structure) took 40 hours of work at the speed of 466 tokens/hour (398 words/hour). This number is comparable to the speed reported by Habash and Roth (2009) (540 tokens/hour). The sentences in their treebank were of the same genre as the data used to train the automatic parsers unlike our case; furthermore, their sentences are much longer than ours (32.0 words/sentence compared to our 7.9 words/sentence). These two issues may explain part of the difference in speed. The end-to-end speed (from raw words to fully corrected trees) is 319 words/hour.

5.7. Inter-Annotator Agreement

To check the consistency of our annotations, we had another person with previous experience in dependency annotation annotate a subset of 100 sentences from this treebank. The second annotator started from the CamelParser output on the same corrected tokenization produced by the first treebank annotator. The inter-annotator agreement scores are 98.7% on POS agreement, 96.1% on label agreement, 90.6% on attachment agreement, and 89.7% on labeled attachment agreement. This is close to the highest average pairwise inter-annotator agreement number reported on the creation of the CATiB Treebank (Habash and Roth, 2009).

6. Results

We present next a comparison between our treebank and the Penn Arabic Treebank, followed by benchmark results of the performance of a state-of-the-art parser on our corpus. We use the abbreviation PATB to refer to facts about the content of the Penn Arabic Treebank, and PATB-CATiB to refer to the CATiB dependency representation of it.

6.1. Comparison with Penn Arabic Treebank

Our corpus is from the travel genre, which has some characteristics that are different from those of the news genre. For example, the average sentence length in MSABTEC is 7.9 words/sentence (9.3 tokens/sentence), as opposed to PATB's average of 32.0 words/sentence (37.6 tokens/sentence). Over 40% of MSABTEC sentences contained a question, while in PATB this percentage did not exceed 2.6%. This is expected as travel corpora are more likely to include questions and answers by travelers.

Moreover, the most frequent words in both corpora vary distinctly. MSABTEC's most frequent verb is يمكن *yumkin* 'can', which is often used when asking for help. In PATB, however, the most common verb is قال *qAl* 'said', which is commonly used for reporting news. In addition, question words such as كم *kam* 'how much', هل *hal* 'do/does', and اين *Âyn* 'where' appear in the set of the most frequent 50 words in MSABTEC, whereas no question words appear in the set respective to PATB. Frequent nouns in MSABTEC include فضل *faĎl* 'favor/please', رقم *raqam* 'number', and غرفة *γurfaħ* 'room'. In PATB, the most frequent nouns include رئيس *raŷiys* 'president', لبنان *lubnAn* 'Lebanon', اليوم *Alyawm* 'today', and المتحدة *Almut∼aHidaħ* 'the united'.

Another phenomenon that differentiates MSABTEC and PATB is the pronoun frequencies. On the one hand, the most frequent pronouns appearing in MSABTEC are +ك *+k*, which is the second person singular pronoun in accusative, and +ي *+y* and +ني *+ny*, which are the first person singular pronouns in genitive and accusative cases, respectively. On the other hand, the most frequent pronouns appearing in PATB are +ه *+h* and +ها *+hA*, which are the masculine and feminine third person singular pronouns, respectively. This leads to the obvious conclusion that MSABTEC mostly contains conversational text that refer to the speaker or the listener, whereas PATB's most dominant style is that of reporting in the third person, which is expected of a news genre corpus.

6.2. Automatic Parsing Quality

We compare the performance of a state-of-the-art parser on our MSABTEC corpus, against its performance on a standard test set from the same corpus it was trained on. The parser we are using is the CamelParser (Shahrour et al., 2016), which was trained and optimized on the PATB-CATiB corpus training set. The results are reported on the PATB-CATiB test set and the entire MSABTEC corpus. The evaluation metrics we are using are Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS), and Label selection, which measure the accuracy of the parser in predicting both the label and the parent, the parent only, and the label only, respectively.

	LAS	**UAS**	**Label**
PATB-CATiB test	83.8%	86.4%	93.2%
MSABTEC	73.5%	77.0%	90.5%

Table 1: The evaluation of the CamelParser prediction on data from PATB-CATiB test and MSABTEC

The error increase in the results of MSABTEC from the results on PATB-CATiB test for the LAS, UAS, and Label selection is 64%, 70% and 39%, respectively. This shows that the genre difference between the training data and the testing data significantly affects the performance of the parser. The previously described characteristics that differ between PATB and MSABTEC (sentence length, prevailing person, and different frequent words) can explain this decline in performance. The large performance drop highlights the need for creating treebanks in less-studied genres to support research on them.

7. Conclusion and Future Work

We presented a small dependency treebank of travel domain sentences in Modern Standard Arabic. The text comes from a translation of the English equivalent sentences in the Basic Traveling Expressions Corpus. The treebank dependency representation is in the style of the

Columbia Arabic Treebank. Our parsing evaluation of the constructed treebank confirms the need for more treebanks in different genres and domains to support research on multi-domain, multi-genre parsers.

In the future, we plan to expand our annotation efforts to other genres and domains as well as to other Arabic dialects. We are also very interested in using the created corpus in improving Arabic syntactic parsing. Since the data we created is small in size compared to the large dominant treebanks, we plan to pursue the genre and domain adaptation research direction. We also plan to make this resource publicly available to support research on Arabic syntactic parsing.

8. Bibliographical References

Abeillé, A. (2012). *Treebanks: Building and using parsed corpora*, volume 20. Springer Science & Business Media.

Alkuhlani, S., Habash, N., and Roth, R. (2013). Automatic morphological enrichment of a morphologically underspecified treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–470, Atlanta, Georgia, June. Association for Computational Linguistics.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Vishnu, S. R., et al. (2017). The Hindi/Urdu treebank project. In *Handbook of Linguistic Annotation*, pages 659–697. Springer.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, May.

Diab, M., Habash, N., Rambow, O., and Roth, R. (2013). LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

Dukes, K. and Buckwalter, T. (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In *Proceedings of the 7th international conference on Informatics and Systems (INFOS 2010)*, Cairo, Egypt.

Eck, M. and Hori, C. (2005). Overview of the IWSLT 2005 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT) 2005*.

Habash, N. and Roth, R. M. (2009). CATiB: The Columbia Arabic Treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224. Association for Computational Linguistics.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., Faraj, R., and Roth, R. (2009). Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Hajič, J., Smrž, O., Zemánek, P., Šnaidauf, J., and Beška, E. (2004). Prague Arabic Dependency Treebank: Development in Data and Tools. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117. ELDA.

Ide, N. and Pustejovsky, J. (2017). *Handbook of Linguistic Annotation*. Springer.

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding semantic annotation to the Penn treebank. In *Proceedings of the human language technology conference*, pages 252–256. San Diego, California.

Lee, D. (2002). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language and Computers*, 42(1):247–292.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006). Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC Vol. 6*.

Maamouri, M., Bies, A., Jin, H., and Buckwalter, T. (2010). The Penn Arabic tree bank. *Computational Approaches to Arabic Script-Based Languages: Current Implementations in Arabic NLP. CSLI NLP Series*.

Maamouri, M., Bies, A., Kulick, S., Tabessi, D., and Krouna, S. (2012). Egyptian Arabic treebank pilot.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn discourse treebank. In *Proceedings of the Language Resources and Evaluation Conference*, Lisbon, Portugal.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Pajas, P. (2008). TrEd: Tree editor. http://ufal.mff.cuni.cz/ pajas/tred.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. M. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Language Re-*

sources and Evaluation Conference (LREC), Reykjavik, Iceland.

Shahrour, A., Khalifa, S., Taji, D., and Habash, N. (2016). CamelParser: A system for Arabic syntactic analysis and morphological disambiguation.

Taji, D., Habash, N., and Zeman, D. (2017). Universal Dependencies for Arabic. *WANLP 2017 (co-located with EACL 2017)*, page 166.

Takezawa, T., Kikui, G., Mizushima, M., and Sumita, E. (2007). Multilingual spoken language corpus development for communication research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Van der Wees, M., Bisazza, A., Weerkamp, W., and Monz, C. (2015). What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *ACL (2)*, pages 560–566.

Webber, B. (2009). Genre distinctions for discourse in the Penn treebank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 674–682. Association for Computational Linguistics.