Constructing a Turkish-English Parallel TreeBank

Olcay Taner Yıldız[†], Ercan Solak[†], Onur Görgün[†,††] and Razieh Ehsani[†]
[†] Işık University, Istanbul, Turkey
[††] Alcatel Lucent Teletaş Telekomünikasyon A.Ş., Istanbul, Turkey
{olcaytaner,ercan,razieh.ehsani}@isikun.edu.tr
onur.gorgun@alcatel-lucent.com

Abstract

In this paper, we report our preliminary efforts in building an English-Turkish parallel treebank corpus for statistical machine translation. In the corpus, we manually generated parallel trees for about 5,000 sentences from Penn Treebank. English sentences in our set have a maximum of 15 tokens, including punctuation. We constrained the translated trees to the reordering of the children and the replacement of the leaf nodes with appropriate glosses. We also report the tools that we built and used in our tree translation task.

1 Introduction

Turkish is an agglutinative and morphologically rich language with a free constituent order. Although statistical NLP research on Turkish has taken significant steps in recent years, much remains to be done. Especially for the annotated corpora, Turkish is still behind similar languages such as Czech, Finnish, or Hungarian. For example, EuroParl corpus (Koehn, 2002), one of the biggest parallel corpora in statistical machine translation, contains 22 languages (but not Turkish). Although there exist some recent works to produce parallel corpora for Turkish-English pair, the produced corpus is only applicable for phrase-based training (Yeniterzi and Oflazer, 2010; El-Kahlout, 2009).

In recent years, many efforts have been made to annotate parallel corpora with syntactic structure to build parallel treebanks. A parallel treebank is a parallel corpus where the sentences in each language are syntactically (if necessary morphologically) annotated, and the sentences and words are aligned. In the parallel treebanks, the syntactic annotation usually follows constituent and/or dependency structure. Well-known parallel treebank efforts are

- Prague Czech-English dependency treebank annotated with dependency structure (Cmejrek et al., 2004)
- English-German parallel treebank, annotated with POS, constituent structures, functional relations, and predicate-argument structures (Cyrus et al., 2003)
- Linköping English-Swedish parallel treebank that contains 1,200 sentences annotated with POS and dependency structures (Ahrenberg, 2007)
- Stockholm multilingual treebank that contains 1,000 sentences in English, German and Swedish annotated with constituent structure (Gustafson-Capkova et al., 2007)

In this study, we report our preliminary efforts in constructing an English-Turkish parallel treebank corpus for statistical machine translation. Our approach converts English parse trees into equivalent Turkish parse trees by applying several transformation heuristics. The main components of our strategy are (i) tree permutation, where we permute the children of a node; and (ii) leaf replacement, where we replace English word token at a leaf node.

This paper is organized as follows: In Section 2, we give the literature review for parallel treebank construction efforts in Turkish. In Section 3, we give a very brief overview on Turkish syntax. We give the details of our corpus construction strategy in Section 4 and explain our transformation heuristics in Section 5. Finally, we conclude in Section 6.

2 Literature Review

Turkish Treebank creation efforts started with the METU-Sabancı dependency Treebank. METU-Sabancı Treebank explicitly represents the head-dependent relations and functional categories. In

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 112–117,
Baltimore, Maryland, USA, June 23-25 2014.

order to adapt the corpus written in 1990's Turkish to further studies, a subset of 7.262 sentences of the corpus was manually annotated morphologically and syntactically (Atalay et al., 2003). METU-Sabancı Treebank is then used in many Turkish NLP studies (Eryigit and Oflazer, 2006; Yuret, 2006; Riedel et al., 2006; Ruket and Baldridge, 2006; Eryigit et al., 2006; Eryigit et al., 2008).

METU-Sabancı Treebank is also subject to transformation efforts from dependency-structure to constituency-structure. Combinatory Categorical Grammar (CCG) is extracted from the METU-Sabancı Treebank with annotation of lexical categories (Cakici, 2005). Sub-lexical units revealing the internal structure of the words are used to generate a Lexical Grammar Formalism (LGF) for Turkish with the help of finite state machines (Cetinoglu and Oflazer, 2006; Cetinoglu and Oflazer, 2009).

Swedish-Turkish parallel treebank is the first parallel Treebank effort for Turkish (Megyesi et al., 2008). The treebank is a balanced syntactically annotated corpus containing both fiction and technical documents. In total, it consists of approximately 160,000 tokens in Swedish and 145,000 in Turkish. Parallel texts are linguistically annotated using different layers from part of speech tags and morphological features to dependency annotation.

English-Swedish-Turkish parallel treebank (Megyesi et al., 2010), mainly the successor of the Swedish-Turkish parallel treebank, consists of approximately 300,000 tokens in Swedish, 160,000 in Turkish and 150,000 in English. The majority of the original text is written in Swedish and translated to Turkish and/or English. For the syntactic description, dependency structure is chosen instead of the constituent structure. All data is automatically annotated with syntactic tags using MaltParser (Nivre et al., 2006a). MaltParser is trained on the Penn Treebank for English, on the Swedish treebank Talbanken05 (Nivre et al., 2006b), and on the METU-Sabancı Turkish Treebank (Atalay et al., 2003), respectively.

ParGram parallel treebank (Sulger et al., 2013) is a joint effort for the construction of a parallel treebank involving ten languages (English, Georgian, German, Hungarian, Indonesian, Norwegian, Polish, Turkish, Urdu, Wolof) from six language families. The treebank is based on deep Lexical-Functional Grammars that were developed within the framework of the Parallel Grammar effort. ParGram treebank allows for the alignment of sentences at several levels: dependency structures, constituency structures and POS information.

3 Turkish syntax

Turkish is an agglutinative language with rich derivational and inflectional morphology through suffixes. Word forms usually have a complex yet fairly regular morphotactics.

Turkish sentences have an unmarked SOV order. However, depending on the discourse, constituents can be scrambled to emphasize, topicalize and focus certain elements. Case markings identify the syntactic functions of the constituents, (Kornfilt, 1997).

4 Corpus construction strategy

In order to constrain the syntactic complexity of the sentences in the corpus, we selected from the Penn Treebank II 9560 trees which contain a maximum of 15 tokens. These include 8660 trees from the training set of the Penn Treebank, 360 trees from its development set and 540 trees from its test set. In the first phase of our work, we translated 4247 trees of the training set and all of those in the development and the test sets.

4.1 Tools

Manual annotation is an error prone task. From simple typos to disagreements among annotators, the range of errors is fairly large. An annotation tool needs to help reduce these errors and help the annotator locate them when they occur. Moreover, the tool needs to present the annotator with a visual tree that is both easy to understand and manipulate for the translation task.

We built a range of custom tools to display, manipulate and save annotated trees in the treebank. The underlying data structure is still textual and uses the standard Treebank II style of syntactic bracketing.

We also implemented a simple statistical helper function within the tool. When translating an English word to a gloss in Turkish, the translator may choose from a list of glosses sorted according their likelihood calculated over their previous uses in similar cases. Thus, as the corpus grows in size, the translators use the leverage of their previous choices.

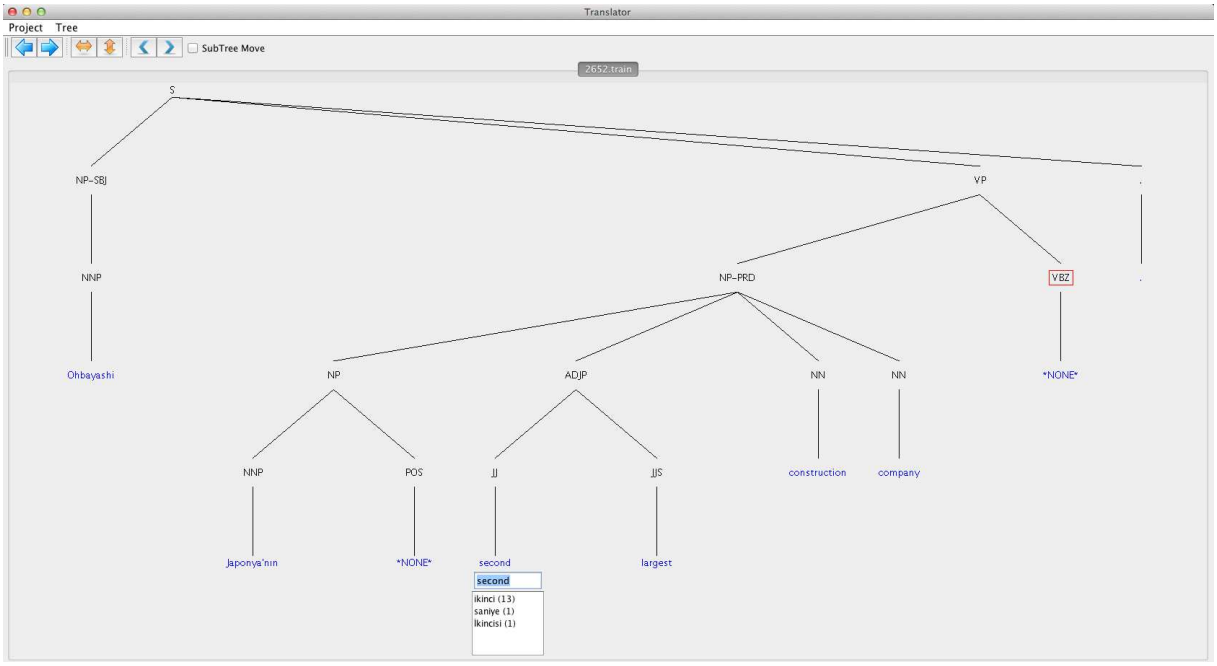

Figure 1: A screenshot of the tree translation tool

Figure 1 shows a screenshot of our tree translation tool.

4.2 Tree permutation

In translating an English syntactic tree, we confine ourselves to two operations. We can permute the children of a node and we can replace the English word token at a leaf node. No other modification of the tree is allowed. In particular, we use the same set of tags and predicate labels in the non-leaf nodes and do not use new tags for the Turkish trees. Adding or deleting nodes are not allowed either.

This might seem like a rather restrictive view of translation. Indeed, it is very easy to construct pairs of translated sentences which involve operations outside our restricted set when transformed into each other.

However, we use the following method to alleviate the restrictions of the small set of operations.

We use the *NONE* tag when we can not use any direct gloss for an English token. In itself, this operation corresponds to effectively mapping an English token to a null token. However, when we use the *NONE* tag, permute the nodes and choose the full inflected forms of the glosses in the Turkish tree, we have a powerful method to convert subtrees to an inflected word. The tree in Figure 2. illustrates this. Note that the POS tag sequence VP-RB-MD-PRP in the Turkish sentence corresponds to the morphological analysis "geç-NEG-FUT-2SG" of the verb "geçmeyeceksin". In general, we try to permute the nodes so as to correspond to the order of inflectional morphemes in the chosen gloss.

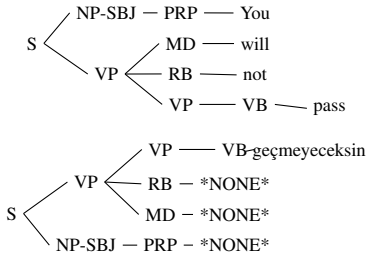

Figure 2: The permutation of the nodes and the replacement of the leaves by the glosses or *NONE*.

5 Transformation heuristics

When we have a sufficiently rich corpus of parallel trees, our next step is to train a SMT learner to imitate the human translator who operates under our restricted set of operations. Naturally, human translators often base their transformation decisions on the whole tree. Still, having a common set of rules and heuristics helps the translators in both consistency and speed. In the following, we illustrate these heuristics.

5.1 Constituent and morpheme order

Majority of unmarked Turkish sentences have the SOV order. When translating English trees, we permute its shallow subtrees to reflect the change of constituent order in Turkish.

Also, the agglutinative suffixes of Turkish words dictate the order when permuting the constituents which correspond to prepositions and particles.

The semantic aspects expressed by prepositions, modals, particles and verb tenses in English in general correspond to specific morphemes attached to the corresponding word stem. For example, "Ali/NNP will/MD sit/VB on/IN a/DT chair/NN" is literally translated as
Ali bir sandalye-ye otur-acak.
Ali a chair-DAT sit-FUT.

If we embed a constituent in the morphemes of a Turkish stem, we replace the English constituent leaf with *NONE*.

In some cases, the personal pronouns acting as subjects are naturally embedded in the verb inflection. In those cases, pronoun in the original tree is replaced with *NONE* and its subtree is moved to after the verb phrase. See Figure 3.

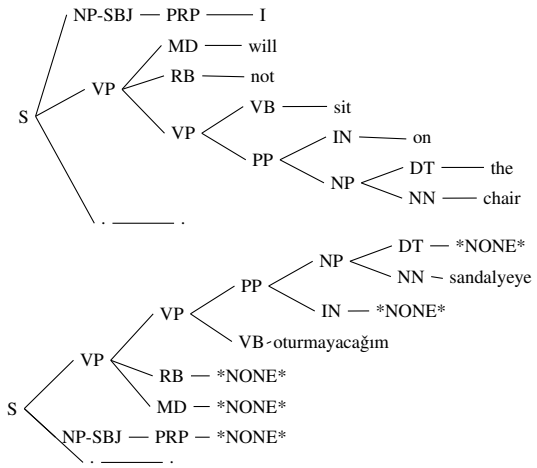

Figure 3: Original and translated trees,
sandalye-ye otur-ma-yacağ-ım
chair-DAT sit-NEG-FUT-1SG

5.2 The determiner "the"

There is no definite article in Turkish corresponding to "the". Depending on the context, "the" is translated either as *NONE* or one of the demonstrative adjectives in Turkish, corresponding to "this" and "that" in English. See Figure 3 .

5.3 Case markers

Turkish, being a fairly scrambling language, uses case markers to denote the syntactic functions of nouns and noun groups. For example, accusative case may be used to mark the direct object of a transitive verb and locative case may be used to mark the head of a prepositional phrase. In translation from English to Turkish, the prepositions are usually replaced with *NONE* and their corresponding case is attached to the nominal head of the phrase. See Figure 4.

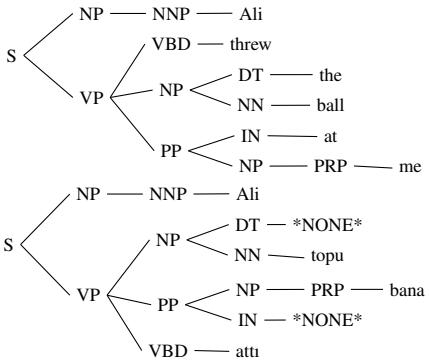

Figure 4: Original and translated trees,
Ali top-u ban-a at-tı
Ali ball-ACC me-DAT throw-PAST-3SG

5.4 Plural in nouns and verb inflection

Number agreement between the verb in the predicate and the subject is somewhat loose in Turkish. We preserved this freedom in translation and chose the number inflection that sounds more natural. Also, plural nouns under NNS tag in the English tree are sometimes translated as singular. In those cases, we kept the original POS tag NNS intact but used the singular gloss. See Figure 5.

5.5 Tense ambiguity

It is in general not possible to find an exact mapping among the tense classes in a pair of languages. When translating the trees, we mapped the English verb tenses to their closest semantic classes in Turkish while trying to keep the overall flow of the Turkish sentence natural. In many cases, we mapped the perfective tense in English to the past tense in Turkish. Similarly, we sometimes mapped the present tense to present continuous. See Figure 5.

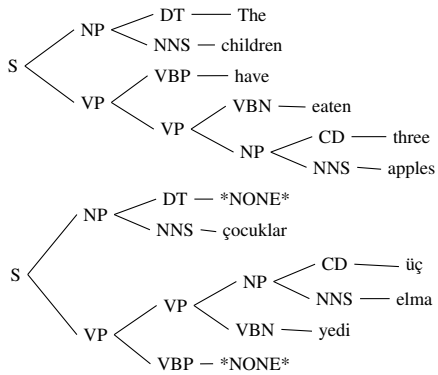

Figure 5: Original and translated trees,
Çocuk-lar üç elma ye-di
Child-PL three apple eat-PAST-3SG

5.6 WH- Questions

Question sentences require special attention during transformation. As opposed to movement in English question sentences, any constituent in Turkish can be questioned by replacing it with an inflected question word. In the Penn Treebank II annotation, the movement leaves a trace and is associated with wh- constituent with a numeric marker. For example, "WHNP-17" and "*T*-17" are associated.

When we translate the tree for a question sentence, we replace the wh- constituent with *NONE* and replace its trace with the appropriate question pronoun in Turkish. See Figure 6.

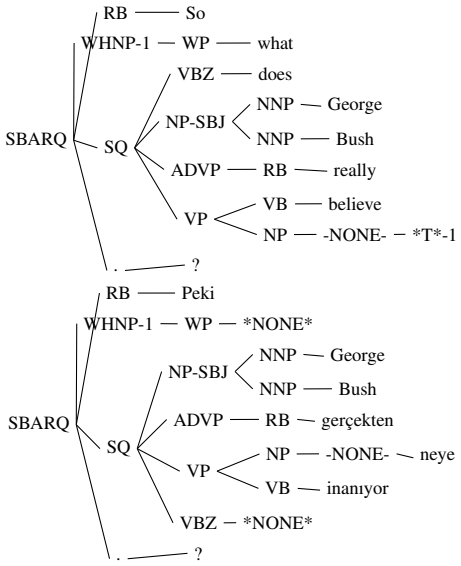

Figure 6: Original and translated trees,
Peki George Bush gerçekten ne-ye inan-ıyor?
So George Bush really what-DAT
believe-PRES-3SG?

5.7 Miscellany

In the translation of nominal clauses, the copula marker "-dIr" corresponding to verb "be" is often dropped.

The proper nouns are translated with their common Turkish gloss if there is one. So, "London" becomes "Londra".

Subordinating conjunctions, marked as "IN" in English sentences, are transformed to *NONE* and the appropriate participle morpheme is appended to the stem in the Turkish translation.

A multiword expression may correspond to a single English word. Conversely, more than one words in English may correspond to a single word in Turkish. In the first case, we use the multiword expression as the gloss. In the latter case, we replace some English words with *NONE*.

6 Conclusion

Parallel treebank construction efforts increased significantly in the recent years. Many parallel treebanks are produced to build statistically strong language models for different languages. In this study, we report our preliminary efforts to build such a parallel corpus for Turkish-English pair. We translated and transformed a subset of parse trees of Penn Treebank to Turkish. We cover more than 50% of all sentences with a maximum length of 15-words including punctuation.

This work constitutes the preliminary step of parallel treebank generation. As a next step, we will focus on morphological analysis and disambiguation of Turkish words. After determining the correct morphological analysis of Turkish words, we will use the parts of these analyses to replace the leaf nodes that we intentionally left as "*NONE*". As a future work, we plan to expand the dataset to include all Penn Treebank sentences.

References

Lars Ahrenberg. 2007. Lines: An english-swedish parallel treebank.

Nart B. Atalay, Kemal Oflazer, and Bilge Say. 2003. The annotation process in the Turkish treebank. In *4th International Workshop on Linguistically Interpreted Corpora*.

Ruken Cakici. 2005. Automatic induction of a ccg grammar for Turkish. In *ACL Student Research Workshop*.

Ozlem Cetinoglu and Kemal Oflazer. 2006. Morphology-syntax interface for Turkish lfg. In *Computational Linguistics and Annual Meeting of the Association*.

Ozlem Cetinoglu and Kemal Oflazer. 2009. Integrating derivational morphology into syntax. In *Recent Advances in Natural Language Processing V*.

Martin Cmejrek, Jan Haji, and Vladislav Kubo. 2004. Prague czech-english dependency treebank: Syntactically annotated resources for machine translation. In *In Proceedings of EAMT 10th Annual Conference*, page 04.

Lea Cyrus, Hendrik Feddes, and Frank Schumacher. 2003. FuSe – a multi-layered parallel treebank. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14–15 November 2003, Växjö, Sweden (TLT 2003)*, volume 9 of *Mathematical Modelling in Physics, Engineering and Cognitive Sciences*, pages 213–216, Växjö. Växjö University Press.

Ilknur D. El-Kahlout. 2009. Statistical machine translation from english to turkish (ph.d. thesis).

Gulsen Eryigit and Kemal Oflazer. 2006. Statistical dependency parsing for Turkish. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Gulsen Eryigit, Esref Adali, and Kemal Oflazer. 2006. Türkçe cümlelerin kural tabanlı bağlılık analizi. In *15th Turkish Symposium on Artificial Intelligence and Neural Networks*.

Gulsen Eryigit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*.

Sofia Gustafson-Capkova, Yvonne Samuelsson, and Martin Volk. 2007. Smultron (version 1.0) - the stockholm multilingual parallel treebank. an english-german-swedish parallel treebank with subsentential alignments.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation.

J. Kornfilt. 1997. *Turkish*. Routledge.

Beáta Megyesi, Bengt Dahlqvist, Eva Pettersson, and Joakim Nivre. 2008. Swedish-turkish parallel treebank. In *LREC*.

Beáta Megyesi, Bengt Dahlqvist, Éva Á. Csató, and Joakim Nivre. 2010. The english-swedish-turkish parallel treebank. In *LREC*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. Maltparser: A data-driven parser-generator for dependency parsing. In *In Proc. of LREC-2006*, pages 2216–2219.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006*, pages 24–26.

S. Riedel, Ruket Cakici, and I. Meza-Ruiz. 2006. Multi-lingual dependency parsing with incremental integer linear programming.

Ruket and Jason Baldridge. 2006. Projective and non-projective Turkish parsing. In *Fifth International Workshop on Treebanks and Linguistic Theories*.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M. Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoglu, I Wayan Arka, and Meladel Mistica. 2013. Pargrambank: The pargram parallel treebank. In *ACL (1)*, pages 550–560.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 454–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deniz Yuret. 2006. Dependency parsing as a classification problem. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.