Parallel Perspectives: ATIS Dependency Treebanks in English and Turkish

Aslı Kuzgun* Olcay Taner Yıldız♡ Mehmet Köse+ Neslihan Cesur*

*Starlang Yazılım Danışmanlık ♡ Özyegin University +Isık University

The ATIS Dataset (Charles T. Hemphill et al., 1990)

- ATIS (Airline Travel Information Systems) is a **domain specific** dataset in English.
- It consists of **audio transcripts** where individuals inquire about **flight details** using automated airline travel systems.
- Main aim of this corpus is to reflect the characteristics of **spoken language** in tourism domain.

The ATIS Dataset in Turkish

- The Turkish ATIS dataset is a **translation** of the original ATIS corpus.
- Translation is done sentence by sentence, **parallel to the original corpus**.
    - Therefore, the two versions are identical to each other in terms of data split.
- The additional information that the ATIS data provides is also protected in the Turkish version.
    - The entities such as airport names, location information is changed re-annotated in Turkish.
    - Intend information did not change.

Outline

- Atis Treebank in Turkish: Overview
- Diversity among the Turkish treebanks:
    - copulars, adjectival -ki, obl vs iobj
- Where does the ATIS stand among other Treebanks?
- Atis in English vs Atis in Turkish
    - What are the different challenges?
- Conclusion

The Atis Treebank in Turkish: Overview

- A **manually annotated** dependency treebank comprising ~46,000 tokens.
- Runs in **parallel** with the English ATIS Treebank.
- Morphological annotation contains 30 universal and 1 **language specific** POS tag.
    - PSOR used for determining the possessor's person in constructs like *'kitabım'* “my book”
- All syntactic relations are **universal** in this treebank.

To tokenize or not tokenize: COP & -*ki* in Turkish

- Atis stands out as a less complex treebank compared to other treebanks in Turkish.
  - It avoids separating **bound forms** on the surface.
  - These forms are represented in morphology.
- This yields a difference with the other treebanks in Turkish.
- The Boun Treebank (Marşan et. al, 2022) and the GB (Çöltekin, 2015) treats these as words.
  - The copular markers -(y)DI, -(y)mIş, -(y)sA , r - (y)ken, -DIr are bound forms, and they are separated.
  - -*ki* is a derivational morpheme that generally turns noun phrases into adjectives, e.g. “evdeki” “the one at home”

Reasons to not tokenize

- **Inconsistencies on copular marking:**
  - In the BOUN treebank, interrogative pronouns such as *kim* and *ne* are **consistently not separated** from COP markers.
    - Nedir ne PRON Ques Case=Nom|Number=Sing|Person=3|PronType=Int 0 root nullcop=3s
  - GB separates these items as ne+dir.

Reasons to not tokenize

- Null cases
    - Sometimes copular is silent, e.g. *Rezan iyi bir pilot*. “Rezan is a good pilot.”
    - The approach that tokenizes COP is not consistent in null cases.

Creates inconsistencies within and across the treebanks.

-*ki*: tokenizing derivational morphemes

- dep:der is a newly introduced dependency relation.
  - This tag is only used in the BOUN treebank to connect the adjectival -ki (PART) to its head noun.

e.g. Mağazalardaki elbiseleri gördüm . \n I saw the dresses at stores
dep:der(ki, Mağazalar)

- GB treats this -ki as ADP and uses case relation.
- ATIS does not separate derivational morphemes **on the surface.**
  - **Nominals with the adjectival -ki are treated as adjectival modifiers.**

IOBJ vs. OBL

- Turkish has two kinds of complements: **direct objects** and **obliques**. (Göksel & Kerslake, 2005)
- Oblique objects refer to individuals or objects indirectly impacted by the verb's action.
  - There is a subgroup of oblique objects carry dative marking.
  - e.g. *Herkes piyanist-**e** bayıl-dı. ‘Everyone adored the pianist.’*
- The dative marked obliques are marked as **iobj** by the BOUN treebank.
- Other treebanks treat them as obliques.

What is different in Turkish Atis in a nutshell

- Copulars are not separated.
- No use of iobj.
- Derivational morpheme -*ki* is not separated.
- All of this information is kept in morphology layer.

Atis in English vs in Turkish

- **Morphological annotation: Combining Resources**
  - Atis in Turkish is annotated by using a **morphological analyzer** (Yıldız, Ercan & Avar, 2019).
  - This step is followed by a fine-tuning performed by human annotators.
  - Atis in English benefited from the **Penn tagset**.
  - Each token matched the most used POS tag in the Penn tagset to be fine tuned by annotators later on.

Conclusion

- Created a parallel treebank from the ATIS data.
  - We hope this new dataset to be useful in the parsing studies in the future in addition to providing a valuable resource for representing linguistic diversity.
- There are different approaches in the Turkish treebank community in one main aspect:
  - What is word? (what should we tokenize?)
- Atis stands in a position for not tokenizing bound morphemes - unless they represent a syntactic dependency- along with the majority of the treebanks in Turkish. (Kenet, FrameNet, Tourism, Penn)

References

Çöltekin, Ç. (2015). A grammar-book treebank of Turkish. In *Proceedings of the 14th Workshop on Treebanks and Linguistic Theories (TLT 14)*.

Göksel, A., & Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.

Marşan, B., Akkurt, S. F., Şen, M., Gürbüz, M., Güngör, O., Özateş, Ş. B., Üsküdarlı, S., Özgür, A., Güngör, T., & Öztürk, B. (2022). Enhancements to the BOUN Treebank Reflecting the Agglutinative Nature of Turkish. *arXiv preprint arXiv:2207.11782*.

Yıldız, O. T., Avar, B., & Ercan, G. (2019). An Open, Extendible, and Fast Turkish Morphological Analyzer. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 1364-1372). INCOMA Ltd.