Demo Proposal: Extensible Integrated Treebank Annotation Environment

Otakar Smrž
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
otakar.smrz@mff.cuni.cz

1 Presentation Overview

The proposed demo would aim to present the technology behind the Prague Arabic Dependency Treebank (Hajič et al., 2004), a project of linguistic annotation having application in many areas of Natural Language Processing.

1.1 TrEd editor

The software environment for maintaining the treebank is TrEd, an editor for tree-like structures developed by Petr Pajas (Hajič et al., 2001). It is a highly customizable and programmable tool providing both the graphical user interface and the application programming interface. TrEd has been used in various annotation projects worldwide, and has been adapted for the Arabic annotation tasks as well.

TrEd integrates all the levels of annotation by enabling the user to invoke macros or external programs of any kind. Thus, given plain text or a document with some markup, e.g. from (Graff et al., 2006), we can run a morphological analyzer or a tagger and prepare files for morphological disambiguation within TrEd. During annotation, we can take great advantage of specific contexts/modes with predefined macro operations, keyboard-shortcuts, and stylesheets for informative display of the data. We further proceed with generating and annotating trees of the analytical syntax, and likewise for the underlying syntax, the so-called tectogrammatics.

The annotation context for morphological disambiguation, the MorphoTrees (Smrž and Pajas, 2004), is of particular interest when processing languages whose scripts allow sequences of lexical words to collapse into a single orthographic word, or whose morphology is rich or complex in some other sense.

In Figure 1, an annotation of the Arabic العربية `AlErbyp` is shown. All the morphological readings of this isolated word are organized into the MorphoTrees hierarchy, which we would explain in detail during the demo. To disambiguate the readings effectively, the annotator can exploit the branching of the hierarchy and take decisions as if in a decision tree to reach the solution in the leaves, or even, can prune the hierarchy with restrictions on the expected morphological properties of the eventual solution. In our example, the selected solution reads العَرَبِيَّةِ *'the-Arabic'*, a feminine singular adjective in definite state and genitive case.

On the level of analytical syntax, in Figure 2, this word is identified as an attribute Atr of the word اللُغَةِ *'the-language'*, and is annotated as its direct grammatical dependent. The dependency approach to syntax is characteristic of the family of Prague Dependency Treebanks, but not of TrEd itself. We have, for instance, implemented contexts for viewing and possibly annotating phrase-structure trees, either in the vertical or the horizontal mode. We would demonstrate this flexibility on data from other treebanks (Bies, 2006). We would also show examples of our Arabic tectogrammatical annotations.

1.2 Other tools

In the rest of the demo, we might briefly present the other tools and resources that we have developed in connection with the treebanking project.

The ElixirFM library implements the Functional Arabic Morphology model (Smrž, 2007). We might explain its extraordinary features, e.g. the design of its lexicon, cf. Figure 3, and discuss other closely related projects, esp. for Arabic (El Dada and Ranta, 2006) and Urdu (Humayoun, 2006).

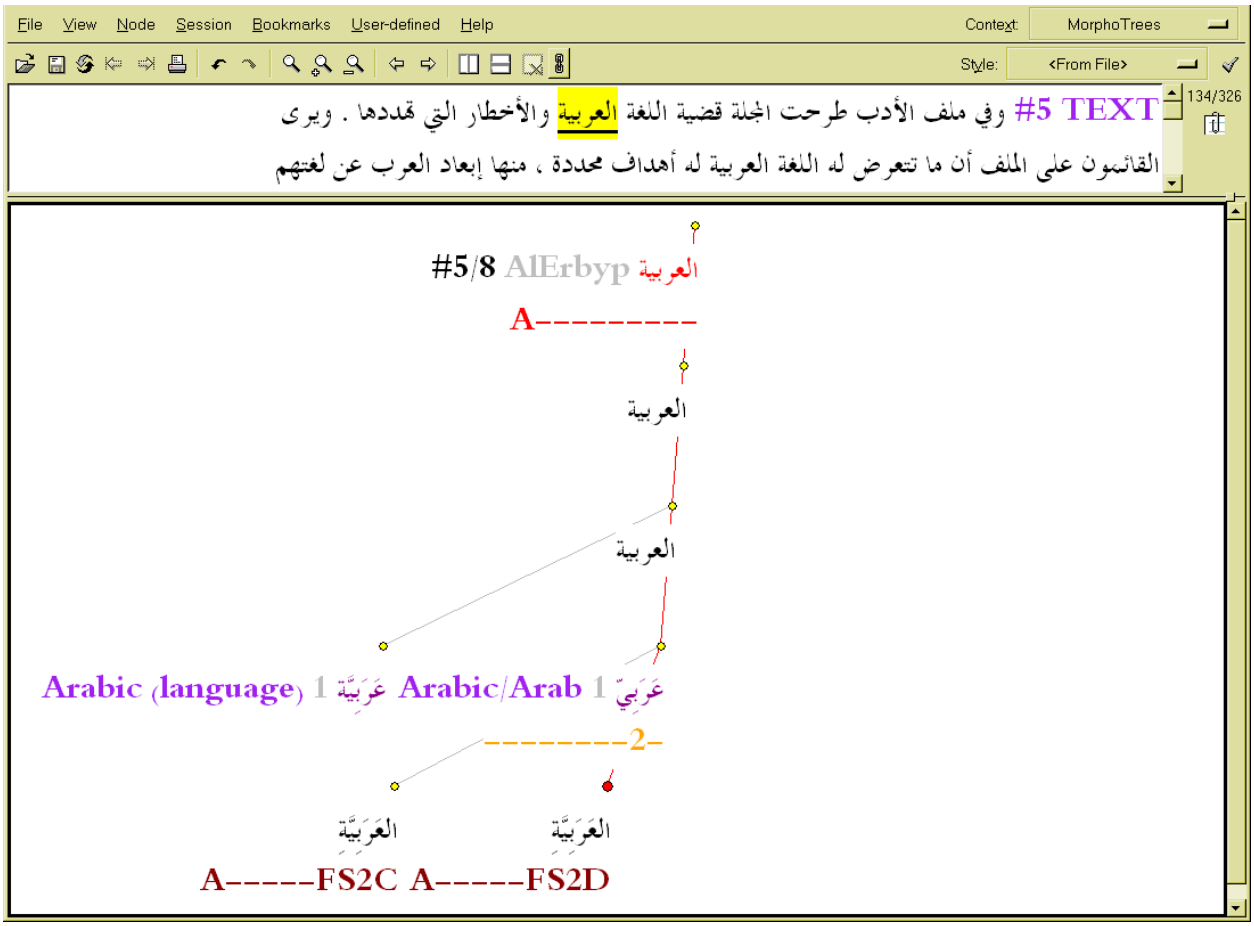

Figure 1: Screenshot of the MorphoTrees annotation in TrEd, with indicated restrictions and the solution.

We could also raise technical issues, cf. TreeX application in Figure 4, or the extensibility of the Encode Arabic project for Farsi, Urdu, etc., cf. below.

2 Further Information

TrEd with its extensions, ElixirFM, and Encode Arabic are open-source software licensed under GNU General Public License. They are available online:

`http://ufal.mff.cuni.cz/~pajas/tred/`

`http://sf.net/projects/elixir-fm/`

`http://sf.net/projects/encode-arabic/`

A video recording of a recent presentation of the Prague Arabic Dependency Treebank can be found on the PADT ++ website:

`http://ufal.mff.cuni.cz/padt/online/2006_12_01_archive.html`

For the demo, we would use our own notebook with all the software and data installed.

This work has been supported by the Ministry of Education of the Czech Republic (MSM00216208-38), by the Grant Agency of Charles University in Prague (UK 373/2005), and by the Grant Agency of the Czech Academy of Sciences (1ET101120413).

References

Ann Bies. 2006. English-Arabic Treebank v 1.0. LDC catalog number LDC2006T10, ISBN 1-58563-387-9.

Ali El Dada and Aarne Ranta. 2006. Open Source Arabic Grammars in Grammatical Framework. In *Proceedings of the Arabic Language Processing Conference (JETALA)*, Rabat, Morocco, June 2006. IERA.

David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2006. Arabic Gigaword Second Edition. LDC catalog number LDC2006T02, 1-58563-371-2.

Jan Hajič, Barbora Hladká, and Petr Pajas. 2001. The Prague Dependency Treebank: Annotation Structure

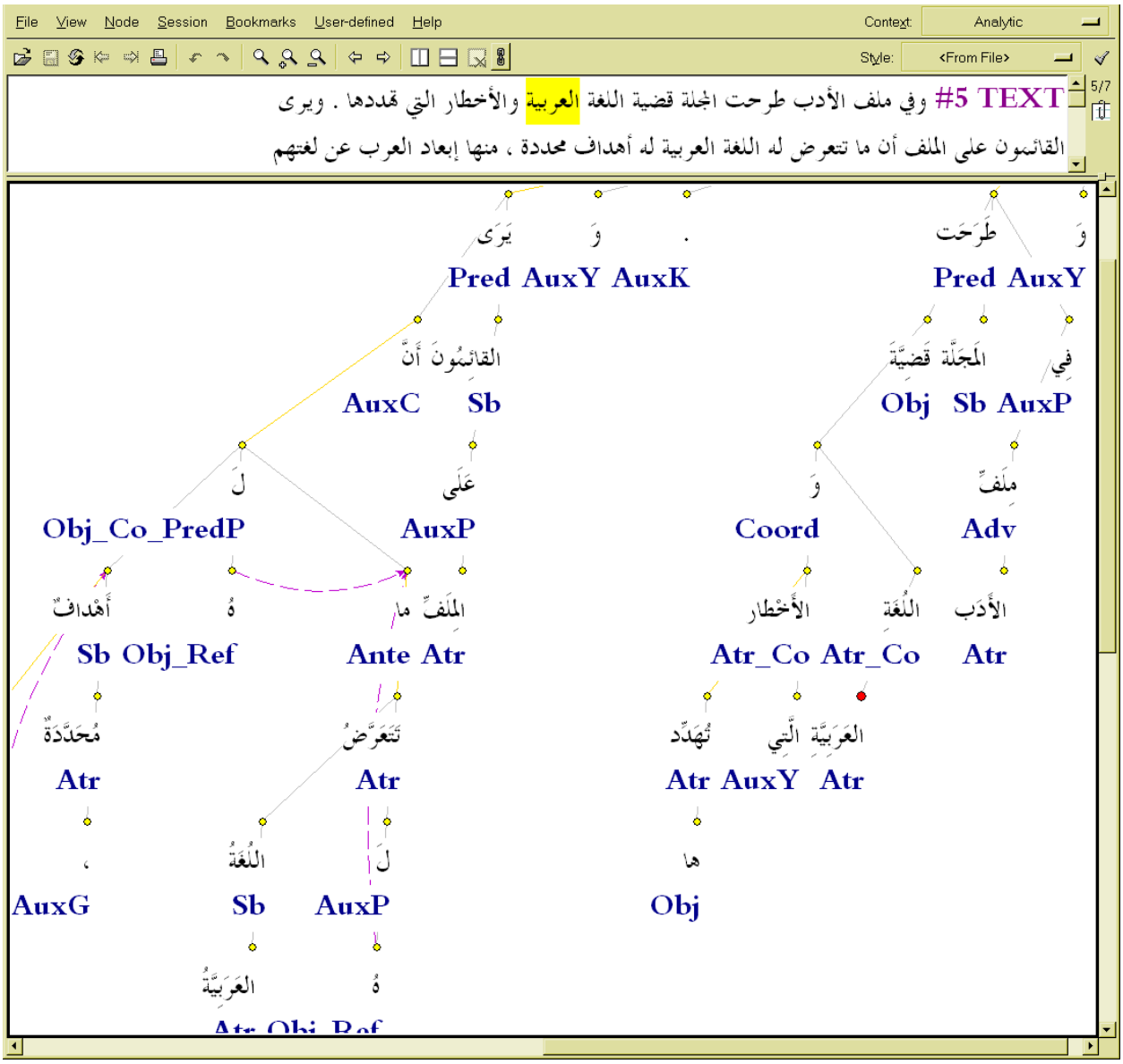

Figure 2: Analytic dependency tree in TrEd, with the highlighted word corresponding to that of Fig. 1.

and Support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 105–114, Philadelphia. University of Pennsylvania.

Jan Hajič, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004. Prague Arabic Dependency Treebank: Development in Data and Tools. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117. ELDA.

Muhammad Humayoun. 2006. Urdu Morphology, Orthography and Lexicon Extraction. Master's thesis, Göteborg University & Chalmers University of Technology, October.

Otakar Smrž and Petr Pajas. 2004. MorphoTrees of Arabic and Their Annotation in the TrEd Environment. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 38–41. ELDA.

Otakar Smrž. 2007. ElixirFM — Implementation of Functional Arabic Morphology. In *ACL 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. ACL, June.

```
|> "k t b" <| [
    FaCaL          'verb'  [ "write", "be destined" ]        'imperf'  FCuL,
    FiCAL          'noun'  [ "book" ]                        'plural'  FuCuL,
    FiCAL |< aT    'noun'  [ "writing" ],
    FiCAL |< aT    'noun'  [ "essay", "piece of writing" ]   'plural'  FiCAL |< At,
    FACiL          'noun'  [ "writer", "author", "clerk" ]   'plural'  FaCaL |< aT
                                                             'plural'  FuCCAL,
    FuCCAL         'noun'  [ "kuttab", "Quran school" ]      'plural'  FaCACIL,
    MaFCaL         'noun'  [ "office", "department" ]        'plural'  MaFACiL,
    MaFCaL |< Iy   'adj'   [ "office" ],
    MaFCaL |< aT   'noun'  [ "library", "bookstore" ]        'plural'  MaFACiL      ]
```

Figure 3: Entries of the ElixirFM lexicon nested under the root *k t b* كتب using morphophonemic templates.

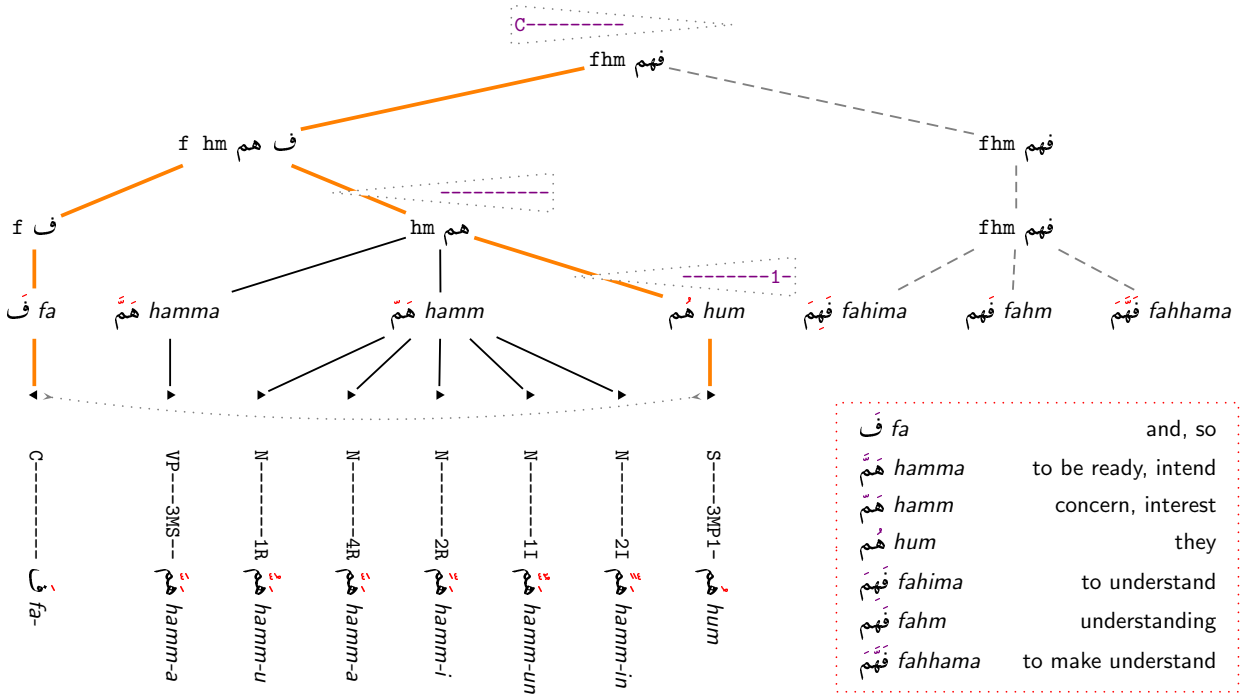

Figure 4: Typesetting MorphoTrees and other data of PADT with TEX/LATEX is easy via the TreeX interface.