# Turkology Annual Online

**Technical Documentation**

Matthias Arnold, Nicolas Bellm, Arina Chitavong, Mateusz Dolata, Anette Frank, Peter Gietz, Jens Hansche, Dustin Heckmann, Christian Roth, Michael Ursinus

September 27, 2010

## 1 Introduction

## 2 Architecture

The architecture is split into three big parts. The first part is the scanning and the OCR with ABBYY FineReader 9.0 Corporate Edition, the second part is the parsing and the storing into a PostgreSQL database and the last part is the web frontend for acessing the database.

First we scanned the printed pages of the 26 volumes of the Turkology Annual. Then we used the Optical Character Recognition (OCR) software ABBYY FineReader 9.0 Corporate Edition to recoginze the scanned text. Therefore we used about 20 scanning languages. Unfortunately, there was no output format which presevered all inforamtion which was gathered by the OCR software. So we used two different output formats to process the scanning output further. The first output format was WordML. This output format preserves hyphenation, formatting and the languages. The second output format was PDF with the original image as overlay. Thus we found out the page number and the exact position in the page of a piece of text. This PDF was converted to plain text with a special patched version of pdftotext which is a part of xpdf to preserve the position information.

For the next step we used the programming language Python 3. In this step we aligned the converted text file and the WordML file. Then we parsed the structure of the entries in the Turkology Annual with the help of LEPL 3, a recursive decent parser for Python 3. We stored the parsing results in an intermediate Python format. The access to the PostgreSQL database happened with SQLAlchemy, a Python SQL Toolkit and Object Relational Mapper (ORM). From the intermediate format we stored the parsed data into this database.

The webfront end is based on the web framework Django. As search engine we used PyLucene which is based on Apache Lucene. So we was able to use the same Object Relational Mapping as we used for the database access in the second step.

# 3 Difficulties

Our approach based on hand-craftet rules and multi-level architecture was not successful with regard to a number of entries from the Turkology Annual. Even though, we were controlling the precision of the developed rules throughout the course of the project, and we tried to improve their performance, we did not reach a point were all the records extracted from the scanned data were correctly parsed. Particularly, introducing general patterns resulted in the increase of false positive, whereas using more precise ones leaded to the decrease of true positives and required the development of further rules with ascending complexity.

In the following, we present several examples of incomplete or incorrect parsing. In order to keep our investigation organised we introduce several *ad-hoc* groups corresponding to the reason for failed analysis. The presented examples are assumed to be representative for the majority of incomplete records within the database. We do not, however, aim to present a full topology of them, i.e., the introduced groups do not need to embrace all incorrect parses. For convenience reasons, we will offer the correct interpretation and the record from the database for every investigated case (cf. Appendix 1).

## 3.1 OCR Problems

This section presents a handful examples of incorrect entries' parsing arising from the inaccuracy of *optical character recognition*. Basically, the multilanguage OCR used within the project does not yield many errors in the recognition of word-like string sequences. Furthermore, such errors does not have much influence on the parsing, so that the database's records are mostly complete in such cases. Conversely, there were lots of mistakes regarding the punctuation marks, which are of great importance to our rule-based paradigm. Most of the entries containing such mistakes could be parsed only partially or not at all.

### 3.1.1 Punctuation

There are lots of problems resulting from bad recognition of punctuation marks.

(1) `1303. Fisher, Alan W. The Ottoman Crimea in the mid-seventeenth century. Some problems and preliminary considerations. In: TA 9.193.215^226.`

In this example the dash between `215` and `226` was recognised by OCR-Machinery as `^`. Therefore the record in the database looks like presented in Figure 1, page 5. It should, however, look like Figure 2, page 5.

### 3.1.2 Numbers

It happens also very often, that the numbers (especially in references) are read as other characters. This is so in the example (2), where `1` has been recognised as `^`, what again

resulted in non-complete record in the database. Moreover, the case presented included a confusing year format and a space before `361-376`, what could also lead to the incorrect parse (cf. Figure 3 and Figure 4, page 5).

(2) `846. Basar, Haşmet Mustafa Kemal Atatürk's thoughts on rural development and co-operative movement. In: ÎFM 38.1^.1980-1981(1982). 361-376.`

## 3.2 Complicated TA-Notation

Another kind of errors are those resulting from confusing, or simply complicated notation used within the Turkology Annual. In most cases errors were caused by unforeseeable specifications of references or comments.

### 3.2.1 References

The syntax of references (both intern and extern) is sometimes difficult to understand and to foresee. An interesting example is included in (3). The resulting records can be seen on page 6.

(3) `745. Delìgönül, Ethem Millî Mücadelede eski Tophaneliler. In: MiKü 3.1.1981.32-35, 2.42-45, 3.44-49. [Die Absolventen der Militärgewerbeschule (Askerî san'at mektebi in Tophane, istanbul) im nationalen Unabhängigkeitskrieg . ]`

The example presented below includes a kind of double reference, whereas the second one, introduced by `Auch erschienen in` is rather an exception than the rule in the notation of Turkology Annual. The algorithm developed in the projec was able to extract the title and the comment, but the reference could not be resolved (c.f. Figures 7 and 8, page 6)

(4) `295. Ivanova, Marija Dumata ile v sistemata na turskite sledlozi, nejnite funkcii i săotvetstvijata ì v bălgarskija ezik. In: BE 31.5.1981.448-451. Auch erschienen in: SåpE 1981.1.54-60. [Das Wort ile im System der türkischen Postpositionen, seine Funktionen und seine Entsprechungen im Bulgarischen.]`

### 3.2.2 Comments

Some entries could not be parsed properly because of the exceptional structure of comments, which normally should be included between [ and ]. The parentheses can be followed by a sequence of characters. But sometimes even this flexible pattern did not capture entries from TA, as the example (5) and the corresponding record (Figure 9 and 10, page 7) show.

(5) `263. Türkiye'de halk ağzından derleme sözlüğü [s. TA 1.203,2.211,5.243]. Weitere Bände: Bd. 10: S-T. Ankara, 1978, S. 3507-4018 (TDK 211/10), Bd. 11: U-Z. Ankara 1979, S. 4019^402 (TDK 211/11), Bd. 12: Ek 1. Ankara, 1982, S. XIV+ 4403-4842(TDK 211/12).`

## 3.3 Parsing

Some of the records within the database are not parsed properly because of the rule-based paradigm underlaying our parsing algorithm. In particular, we try to match the entries from TA to a number of patterns, from the specific ones to the more general ones. Sometimes, the input string is not matched by any of the patterns. In some other cases it is not matched by the appropriate specific pattern, but it is captured by a general one. Both cases are discussed below.

### 3.3.1 Too general parse

Example (6) presents a formally properly constructed entry for a collection including the information on editors (*Richard L. Chambers—Günay Kut (Alpay) ed.*). The latter does not, however, exhibit a standard form. Firstly, it includes two editors. Secondly, the name of the first person includes the initial of the middle name followed by a period, which is at the same time a quite important idiosyncratic symbol, terminating, e.g. titles or comments. And, last but not least, this string includes parentheses, which is also unusual. For the resulting record, cf., Figure 11 and 12, page 7.

(6) `316. Contemporary Turkish short stories. An intermediate reader. Richard L. Chambers|Günay Kut (Alpay) ed. Minneapolis|Chicago, 1977, XI + 159 S. (Middle Eastern Languages and Linguistics, 3).`

### 3.3.2 No parse possible

Some non-standard entries resulted in no parse at all. This can have several reasons, and mostly, for every case special pattern would be needed. Another approach could be based on statistical methods, that use the properly recognised entries as training set. We assume, that such an approach would significantly decrease the number of non-parsed recods. The example below includes an entry that could not be parsed properly by our algorithm. It is so probably due to the many numbers occurring in the title. And due to the small mistake in the reference (there is a space between `1985.` and `113-130.`). We present the proper parse for this example in Figure 13, page 8.

(7) `586. Gallotta, Aldo Venise et l'Empire ottoman, de la paix du 25 janvier 1479 à la mort de Mahomet II, 3 mai 1481. In: ROMM 39.1985. 113-130.`

## 4 Conclusions

## Appendix 1

| | |
|---|---|
| ID: | 20924 |
| Type: | article |
| Autor: | Fisher, Alan W. |
| Title: | The Ottoman Crimea in the mid-seventeenth century. Some problems and preliminary considerations. |

Figure 1: Database record for the example in (1)

| | |
|---|---|
| ID: | 20924 |
| Type: | article |
| Autor: | Fisher, Alan W. |
| Title: | The Ottoman Crimea in the mid-seventeenth century. Some problems and preliminary considerations. |
| In: | TA 9.193.215-226. |

Figure 2: Correct record for the example in (1)

| | |
|---|---|
| ID: | 20227 |
| Type: | article |
| Autor: | Basar, Haşmet |
| Title: | Mustafa Kemal Atatürk's thoughts on rural development and co-operative movement. |

Figure 3: Database record for the example in (2)

| | |
|---|---|
| ID: | 20227 |
| Type: | article |
| Autor: | Basar, Haşmet |
| Title: | Mustafa Kemal Atatürk's thoughts on rural development and co-operative |
| In: | ÎFM 38.11.1980-1981(1982).361-376. |

Figure 4: Correct record for the example in (2)

| | |
|---|---|
| ID: | 20106 |
| Type: | article |
| Autor: | Delìgönül, Ethem |
| Title: | Millî Mücadelede eski Tophaneliler. |
| Comment: | [Die Absolventen der Militärgewerbeschule (Askerî san'at mektebi in Tophane, istanbul) im nationalen Unabhängigkeitskrieg .] |

Figure 5: Database record for the example in (3)

| | |
|---|---|
| ID: | 20106 |
| Type: | article |
| Autor: | Delìgönül, Ethem |
| Title: | Millî Mücadelede eski Tophaneliler. |
| Comment: | [Die Absolventen der Militärgewerbeschule (Askerî san'at mektebi in Tophane, istanbul) im nationalen Unabhängigkeitskrieg .] |
| In: | MiKü 3.1.1981.32-35, 2.42-45, 3.44-49 |

Figure 6: Correct record for the example in (3)

| | |
|---|---|
| ID: | 19549 |
| Type: | article |
| Autor: | Ivanova, Marija |
| Title: | Dumata ile v sistemata na turskite sledlozi, nejnite funkcii i săotvetstvijata ì v bălgarskija ezik. |
| Comment: | [Das Wort ile im System der türkischen Postpositionen, seine Funktionen und seine Entsprechungen im Bulgarischen.] |

Figure 7: Database record for the example in (4)

| | |
|---|---|
| ID: | 19549 |
| Type: | article |
| Autor: | Ivanova, Marija |
| Title: | Dumata ile v sistemata na turskite sledlozi, nejnite funkcii i săotvetstvijata ì v bălgarskija ezik. |
| Comment: | [Das Wort ile im System der türkischen Postpositionen, seine Funktionen und seine Entsprechungen im Bulgarischen.] |
| In: | BE 31.5.1981.448-451.; SåpE 1981.1.54-60. |

Figure 8: Correct record for the example in (4)

| | |
|---|---|
| ID: | 19509 |
| Type: | collection |
| Title: | Türkiye'de halk ağzından derleme sözlüğü [s. TA 1.203,2.211,5.243]. Weitere Bände: Bd. 10: S-T. |
| Location: | Ankara |
| Year: | 1978 |
| Comment: | 4019ˆ402 (TDK 211/11), Bd. 12: Ek 1. Ankara, 1982, S. XIV+ 4403-4842(TDK 211/12). |

Figure 9: Database record for the example in (5)

| | |
|---|---|
| ID: | 19509 |
| Type: | collection |
| Title: | Türkiye'de halk ağzından derleme sözlüğü |
| Location: | Ankara |
| Year: | 1978 |
| Comment: | [s. TA 1.203,2.211,5.243]. Weitere Bände: Bd. 10: S-T. 4019ˆ402 (TDK 211/11), Bd. 12: Ek 1. Ankara, 1982, S. XIV+ 4403-4842(TDK 211/12). |

Figure 10: Correct record for the example in (5)

| | |
|---|---|
| ID: | 5897 |
| Type: | collection |
| Title: | Contemporary Turkish short stories. An intermediate reader. Richard L. Chambers—Günay Kut (Alpay) ed. |
| Location: | Minneapolis—Chicago |
| Year: | 1977 |
| Comment: | (Middle Eastern Languages and Linguistics, 3). |

Figure 11: Database record for the example in (6)

| | |
|---|---|
| ID: | 5897 |
| Type: | collection |
| Title: | Contemporary Turkish short stories. An intermediate reader. |
| Location: | Minneapolis—Chicago |
| Year: | 1977 |
| Comment: | (Middle Eastern Languages and Linguistics, 3). |
| Editor: | Richard L. Chambers; Günay Kut (Alpay) |

Figure 12: Correct record for the example in (6)

| | |
|---|---|
| ID: | 38742 |
| Type: | article |
| Autoren: | Galotta, Aldo |
| Title: | Venise et l'Empire ottoman, de la paix du 25 janvier 1479 à la mort de Mahomet II, 3 mai 1481. |
| In: | ROMM 39.1985. 113-130. |

Figure 13: Correct record for the example in (7)