The Historical Arabic Dictionary Corpus and its Suitability for a Grammaticalization Approach

Almoataz B. Al-Said
Cairo University
Lucía Medea-García
CLT / Val.Es.Co (Autonomous University of Barcelona)

Grammar and Corpora, 25-27 June, University of Warsaw
Outline

1. Introduction
2. Overview
3. Characteristics of the HADC
4. The use of the HADC for a Grammaticalization Approach
   a. Advantages
   b. Drawbacks
5. Relevance
6. Conclusions
1. Introduction

Grammaticalization: “The change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions” (Hopper & Traugott: 2003).
Grammaticalization: “The change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions” (Hopper & Traugott: 2003).

- Very important for the development of the studies on Historical Syntax
- Analyzed in languages from all linguistic families
- Corpus linguistics has played a major role in its progress and consolidation (Mair 2011)
1. Introduction

Grammaticalization: “The change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions” (Hopper & Traugott: 2003).

- Very important for the development of the studies on Historical Syntax
- Analyzed in languages from all linguistic families
- Corpus linguistics has played a major role in its progress and consolidation (Mair 2011)

Barely applied to the Arabic language
2. Overview

- Difficulties:
2. Overview

Difficulties:

1. Complex sociolinguistic situation of diglossia (Ferguson 1959)
2. Sacred nature of the Classical Arabic
3. Lack of resources
4. Paucity of research
2. Overview

- Insufficiency of historical corpora:
  - Arabic Corpus Tool (Parkinson et al.)
  - King Saud University Corpus of Classical Arabic (Althubaity et al. 2013)
  - The Quranic Arabic Corpus (University of Leeds)
2. Overview

- Insufficiency of historical corpora:
  - Arabic Corpus Tool (Parkinson et al.)
  - King Saud University Corpus of Classical Arabic (Althubaity et al. 2013)
  - The Quranic Arabic Corpus (University of Leeds)

The Historical Arabic Dictionary Corpus (Al-Said 2011)
3. Characteristics of the HADC

- Developed by Dr. Al-Said (2011) at Cairo University
- Originally designed to build a historical dictionary
- Texts in Classical Arabic and Modern Standard Arabic
- More than 116 millions of words
3. Characteristics of the HADC

- Developed by Dr. Al-Said (2011) at Cairo University
- Originally designed to build a historical dictionary
- Texts in Classical Arabic and Modern Standard Arabic
- More than 116 millions of words
- From the 2nd up to the 20th century
- At least 1 million words per century
- The only corpus that takes into account so many different variables
3. Characteristics of the HADC

- Developed by Dr. Al-Said (2011) at Cairo University
- Originally designed to build a historical dictionary
- Texts in Classical Arabic and Modern Standard Arabic
- More than 116 millions of words
- From the 2nd up to the 20th century
- At least 1 million words per century
- The only corpus that takes into account so many different variables
- Unpublished
3. Characteristics of the HADC

Variables taken into account in the HADC:

1. **Historical (3):**
   - Age
   - Century
   - Year of the author’s death
3. Characteristics of the HADC

- Variables taken into account in the HADC:

1. **Historical (3):**
   - Age
   - Century
   - Year of the author’s death

2. **Geographical (7):**
   - Al-Andalus
   - Maghreb and Sicily
   - Nile Valley
   - The Levant
   - Arabian Peninsula
   - Mesopotamia
   - Persia and Transoxiana
3. Characteristics of the HADC

3. Literary genres and text types (15):
   - Poetry
   - Quran
   - Literary prose
   - Hadiths
   - History and Genealogy
   - Religions and Doctrines
   - Encyclopedias and Dictionaries
   - Journalistic texts
   - Geography and Travel literature
3. Characteristics of the HADC

3. Literary genres and text types (15):
   - Poetry
   - Quran
   - Literary prose
   - Hadiths
   - History and Genealogy
   - Religions and Doctrines
   - Encyclopedias and Dictionaries
   - Journalistic texts
   - Geography and Travel literature

4. Author
3. Characteristics of the HADC

3. Literary genres and text types (15):
   - Poetry
   - Quran
   - Literary prose
   - Hadiths
   - History and Genealogy
   - Religions and Doctrines
   - Encyclopedias and Dictionaries
   - Journalistic texts
   - Geography and Travel literature

4. Author

5. Number of words
   The texts contain the information about the number of words they have.
3. Characteristics of the HADC
3. Characteristics of the HADC
3. Characteristics of the HADC
3. Characteristics of the HADC
3. Characteristics of the HADC

- العصر الجاهلي
- العصر الإسلامي
- العصر العباسي
- العصر الوسيط
- العصر الحديث

- كتلة رقية - 0100
- شعر ابن ميادة - 0149
- نثر مباشر - 0150
- شعر بشر بن برد - 0167
- جبرت بن علي الأندلي - 0169
- سهيل بن عبد العزيز - 0170
- شعر عبد الرحمن الجاهلي - 0177
- شعر البيضاء الجاهلي - 0173
- الأصحاب للجاهلي - 0178
- الكتاتب - 0180
- شعر عبد الرحمن - 1801
- شعر مروان بن أبي حفص - 0182
- شعر أبو الفضل بن أحمد - 0192
- الأشاعر للمجاهلي - 0195
- شعر أبي الشيخ محمد - 0196
- شعر أبو نواس - 0198
- شعر أبو الفتح - 0199
- شعر أبي الجعفي - 0200
- الأشاعر - 0204
- نثر ولمان - 0205
- شعر لشافعي - 0206
3. Characteristics of the HADC

In this section, we will examine the characteristics of the HADC, which is a significant body in the Arab world. The HADC is a council that represents the interests of Arab states in international affairs. It is composed of representatives from Arab countries and plays a key role in promoting Arab unity and cooperation. The HADC is known for its strong stance on issues related to the Arab-Israeli conflict and the rights of the Palestinian people.

The HADC is headquartered in Damascus, Syria, and its members include all Arab League countries except Jordan. The council operates through periodic meetings and working groups to discuss and address various issues affecting the Arab world.

The HADC is recognized as a platform for Arab states to coordinate their foreign policies and to present a united front in international forums. It also serves as a forum for member states to exchange views and ideas on regional and global issues.

The council is instrumental in raising awareness of Arab causes and advocating for the rights of the Palestinian people. It has played a significant role in supporting the Palestinian struggle for independence and self-determination.

In conclusion, the HADC is a vital body in the Arab world, serving as a strong voice for Arab interests and unity. Its role in addressing regional and global issues is crucial for the development and progress of the Arab world.
3. Characteristics of the HADC

3. Characteristics of the HADC
3. Characteristics of the HADC

<table>
<thead>
<tr>
<th>عدد الكلمات</th>
<th>اسم الراوي</th>
<th>بث (ت)</th>
<th>عنوان الوثيقة</th>
<th>م</th>
</tr>
</thead>
<tbody>
<tr>
<td>75,091</td>
<td>الشام</td>
<td>244</td>
<td>الزبير الصابئ</td>
<td>1</td>
</tr>
<tr>
<td>12,622</td>
<td>الرافدا</td>
<td>256</td>
<td>في الفلسفة الأولى</td>
<td>2</td>
</tr>
<tr>
<td>1,163,593</td>
<td>فارس</td>
<td>311</td>
<td>الحاوي في الطب</td>
<td>3</td>
</tr>
<tr>
<td>14,957</td>
<td>فارس</td>
<td>339</td>
<td>الألفاظ المستعملة في المنطق</td>
<td>4</td>
</tr>
<tr>
<td>9,713</td>
<td>فارس</td>
<td>339</td>
<td>تحصيل السعادة</td>
<td>5</td>
</tr>
<tr>
<td>45,202</td>
<td>فارس</td>
<td>339</td>
<td>رسالة الحروف</td>
<td>6</td>
</tr>
<tr>
<td>7,813</td>
<td>فارس</td>
<td>339</td>
<td>في المنطق</td>
<td>7</td>
</tr>
<tr>
<td>119,223</td>
<td>الرافدا</td>
<td>385</td>
<td>الفهرست</td>
<td>8</td>
</tr>
<tr>
<td>27,771</td>
<td>فارس</td>
<td>387</td>
<td>مفاتيح العلوم</td>
<td>9</td>
</tr>
</tbody>
</table>
4. The use of the HADC for a grammaticalization approach

Methodology:
4. The use of the HADC for a grammaticalization approach

Methodology:

1. Historical periods:
   1. 7th century. The Quran.
   2. 9th century. Beginnings of the Abbasid Caliphate.
   5. 19th century. Al-Nahda.
4. The use of the HADC for a...

2. Discourse tradition:

• Historiographic texts:
  History
  Genealogy
  Travel literature
  Geography (excluding scientific treatises)
4. The use of the HADC for a...

3. Geographical origin of the author:

We adopt a pan-Arabist perspective, with the variable of the origin of the author controlled.

It helps us to select texts in particular historical periods.
4. The use of the HADC for a...

The control of these variables

Taking into account the variable of the language

Selection of tokens to analyze

Analysis datasheet (syntactic, semantic and pragmatic information)
Y construyó el sultán un puente sobre el Tigris y cruzó al lado oriental de manera que tuvo lugar la lucha en ambos lados.

Muy interesante, porque se queda solo con el significado de la consecuencia, no de la intensificación. ¡Ya en el XIV!

En principio no podría ser final ('para') porque el verbo tras hatta está en pasado, y porque la lectura consecutiva es posible.
4. The use of the HADC for a…

The control of these variables

Taking into account the variable of the language

Selection of tokens to analyze

Analysis datasheet
(syntactic, semantic and pragmatic information)

Study of the evolutionary process of a particle throughout 1400 years
4. The use of the HADC for a...

The control of these variables

Selection of tokens to analyze

Taking into account the variable of the language

Analysis datasheet (syntactic, semantic and pragmatic information)

Study of the evolutionary process of a particle throughout 1400 years
4a. Advantages

1. It is the only historical corpus we have of the Arabic language

2. It is the only one that takes into account so many variables

3. It does not have any historical gap

4. It contains texts from the 2nd century

5. It can be modified and adapted to specific necessities by the author

6. It offers all the context of the word or construction under study
4b. Drawbacks

1. Lack of accessibility

2. Two different programs to obtain the information needed:
   • Shameela (exclusively in Arabic)
   • Nooj

3. Variables not accessible automatically

4. Not online

5. Absence of some important texts

6. Total absence of women’s contributions (Kahala 1985)
Corpora are a fundamental tool for diachronic linguistics

Corpora and the use of corpus-linguistic methods enrich the empirical basis of grammaticalization theory and even help to develop its theoretical foundations (Mair 2013)
5. Relevance

Corpora are a fundamental tool for diachronic linguistics

1. In the Semitic family, Arabic is one of the most advantageous languages for a diachronic study:
5. Relevance

Corpora are a fundamental tool for diachronic linguistics

1. In the Semitic family, Arabic is one of the most advantageous languages for a diachronic study:
   
   • It includes a lot of different dialects
   • It experiences and has experienced language contact with languages from very different linguistic families
   • It has written texts uninterruptedly from the 1st century
   • It boasts one of the most ancient and prolific grammatical traditions (Versteegh 1997)
5. Relevance

2. The Arabic language would benefit greatly from a grammaticalization approach:
5. Relevance

2. The Arabic language would benefit greatly from a grammaticalization approach:

- Offering different methodologies for its study
- Helping to give prestige to the dialects and to include them in the linguistic studies in a more systematic way
- Favoring the development of resources for its study
- Getting rid of some obsolete postulates
5. Relevance

3. Grammaticalization studies should take into account the evolution of the Arabic language:
3. Grammaticalization studies should take into account the evolution of the Arabic language:

- Huge variety of dialects
- Many information about the processes of linguistic contact
- Exploring the possible particularities in the grammaticalization processes of Arabic (Bisang 2008)
- It could help to improve certain methodological approaches (Heine 2000; Owens 2006)
6. Conclusions

- Presenting the HADC, which is the only historical corpus available for the Arabic language that allows to carry out such a detailed research.

- Proposing a methodology to analyze grammaticalization processes in Arabic in a more rigorous and systematic way.

- Vindicating the need to include the Arabic language in diachronic studies.
References

References

References

Thank you!

moataz@cu.edu.eg
luciamedeea.garcia@uab.cat