Expressing referentiality in languages with and without articles

Laura Becker

November 16, 2018,
Grammar and Corpora
(Paris Diderot)
In a nutshell

Background

Although definiteness and referentiality have received a lot of attention, there are still many open questions.

**Question 1** How many (and which) referential values need to be distinguished?

**Question 2** What is the role of articles?

How important are other properties of the expressions to determine its referential value?

Today, I will address these 2 questions.

- using parallel texts (movie subtitles), referential expressions across languages can be compared and semantics can be controlled for
- I will compare German & Spanish (def. and indef. articles) vs. Macedonian (def. article) vs. Russian (no articles)
Referentiality

“Traditional” distinctions of referential values

- commonly distinguished values
  - (in)definite, (non)specific, generic
- a more fine-grained distinction
  - givenness hierarchy (Gundel, Hedberg, and Zacharski 1993)
  - reference hierarchy (Dryer 2014)

- ...

The present distinctions

(cf. Hawkins 1978; Ariel 1988; Himmelmann 1997; Dryer 2014; Becker 2018)

- deictic (identifiable by perception)
- anaphoric (identifiable by previous mention)
- (situationally) unique (identifiable by situational uniqueness)
- bridging (identifiable by unique link to other referent)
- establishing (not yet identifiable, but soon to be)
- specific (not identifiable, but a particular referent)
- nonspecific (not identifiable, no particular referent)
**Deictic referent**

(1) (Pointing to birthday presents) Aren’t **they** wonderful, darling?

de Sind **die** nicht wunderbar, Schätzchen?
are they not wonderful darling

sp ¿No son maravillosos, corazón?
not are wonderful darling

mk Не ли се прекрасни, душо?
not if are wonderful darling

ru Разве **они** не чудесны, дорогой?
really they not wonderful darling
Examples

Anaphoric referent

(2) He’ll be famous.

dE Er wird berühmt werden.
he will famous become

sp Este niño será famoso.
this child will be famous

mk Ова дете ќе биде познато.
this child will be known

ru Он знаменит.
he famous
Examples

Bridging referent

(3) I’m told it’s the latest fashion in London. Well, **women in London** must’ve learned not to breathe!

de  **Die Frauen** müssen gelernt haben, nicht zu atmen.

the women must learned have not to breathe

sp  **¡Las londinenses** deben haber aprendido a no respirar!

the.f londoners must have learned to not breathe

mk  **Жени-те во Лондон** веројатно научиле да не дишат.

women-the in London really learned comp not breathe

ru  Наверное, **лондонские модницы** научились обходиться без really london.adj stylish.people learned exist without воздуха.

air
Examples

Situationally unique referent

(4) Rouse the captain immediately.

de Weckt den Captain.
  wake.up the captain

sp Levanta al capitán.
  rouse PREP.the captain

mk Веднаш викнете го капитан-от!
  immediately call 3sg captain-the

ru Доложить капитану.
  report captain.DAT
Establishing referent

(5) Are **the rumors** true, Albus?

de Darf man **den Gerüchten** trauen, Albus?

sp ¿Son ciertos **los rumores**, Albus?

mk Вистинити се **озборувања-та**, Албус?

ru **Слухи** верны, Альбус?
Specific (indefinite) referent

(6) If you have a few moments, Mr. Cobb has a job offer he’d like to discuss with you.

de Wenn Sie einen Moment Zeit haben, würde Mr. Cobb gerne ein Jobangebot mit Ihnen besprechen.

sp Si tienes un momento, el Sr. Cobb quiere ofrecer-te un trabajo.

mk Ако имаш некоја минутка, г-дин Cobb има бизнис понуда за тебе.

ru Если у тебя есть время, у мистера Кобба к тебе деловое предложение.
Examples

Nonspecific referent

(7) A work placement?

de Geht’s um ein Praktikum?
goes.it about a internship

sp Un internado?
a internship

mk Стажиранье?
internship

ru Наверное, стажировка?
really, internship
Movies *Inception, Harry Potter 1, Pirates of the Caribbean 1, The Lion King, Lord of the Rings 1*

Data set 379 referential expression for each German, Spanish, Macedonian, and Russian

Annotation

- **referentiality** deictic, anaph, sit.unique, estab, bridging, specific, nonspec
- **syntactic function** subj, obj, obl, other
- **semantic type** human, concrete, abstract, place
- **expression type** np, pro, drop
- **article** def, indef, no
- **possessive** yes, no
- **demonstrative** yes, no
- **adjective** yes, no
- **other attribute** yes, no
- **number** sg, pl
Results
Major clusters of referential values

The clustering method used: correlation-based distance (1-cor) for gram.function, art, poss, dem, adj, other.attr, expression, number

Three major referential values can be distinguished:
- Activated definite (anaphoric, deictic)
- Definite (establishing, bridging, situationally unique)
- Indefinite (specific, nonspecific)
Major clusters of referential values

Three major referential values can be distinguished:

- **activated definite** (anaphoric, deictic)
- **definite** (establishing, bridging, situationally unique)
- **indefinite** (specific, nonspecific)
Major clusters for each language

Clustering referential values (German)

Clustering referential values (Spanish)

Clustering referential values (Macedonian)

Clustering referential values (Russian)
Distribution of possessives

possessive across referential values

count

delictic anaph

bridging sit.unique estab

specific nonspec
Distribution of demonstratives

demonstrative across referential values

count

- delictic
- anaph

- bridging
- sit.unique
- estab

- specific
- nonspec
Distribution of expression type

Results

Clustering referential values
Distribution of other attributes

other attributes across referential values

count

specific	nonspec
Modelling referentiality with random forests

Random forests are ensembles containing a large numbers of classification trees.

- they use recursive partitioning and random variable selection and explore the correlations between different variables is in different subsets of the data
- forests are relatively stable with respect to collinearity
- they are especially useful for this kind of data (unbalanced, many correlated predictors)

(Strobl, Malley, and Tutz 2009)

The model

\[
\text{ref} \sim \text{gram.function} + \text{semantics} + \text{expression.type} + \\
\text{art} + \text{poss} + \text{dem} + \text{adj} + \text{other.attr} + \text{number}
\]

with ref having the three major values: act.def, def, indef
Variable importance

German

Spanish
Variable importance

Macedonian

Russian

Results

Modelling the expression of referentiality
# Accuracy of the Models

How well the fitted models perform can be assessed by a confusion matrix.

<table>
<thead>
<tr>
<th>Language</th>
<th>German</th>
<th>Spanish</th>
<th>Macedonian</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicted</strong></td>
<td>act-def</td>
<td>def</td>
<td>indef</td>
<td>act-def</td>
</tr>
<tr>
<td>act-def</td>
<td>90</td>
<td>14</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>def</td>
<td>36</td>
<td>118</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>indef</td>
<td>8</td>
<td>8</td>
<td>100</td>
<td>7</td>
</tr>
</tbody>
</table>

**Accuracy**:
- **German**: 0.8127
- **Spanish**: 0.8021
- **Macedonian**: 0.7203
- **Russian**: 0.6728

90% CI:
- German: (0.7697, 0.8507)
- Spanish: (0.7584, 0.841)
- Macedonian: (0.6722, 0.7649)
- Russian: (0.6231, 0.7199)

No Information Rate:
- 0.3694

---

**Results**

Modelling the expression of referentiality

2018
Accuracy of the Models (without the article)

Without the article as predictor, the accuracy of the models for German, Spanish, and Macedonian drop to the one of the Russian model.

<table>
<thead>
<tr>
<th>Language</th>
<th>observed</th>
<th>predicted</th>
<th>act-def</th>
<th>def</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>act-def</td>
<td></td>
<td>90</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>def</td>
<td></td>
<td>28</td>
<td>92</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>indef</td>
<td></td>
<td>16</td>
<td>34</td>
<td>80</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.6913</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% CI</td>
<td></td>
<td></td>
<td>(0.6421, 0.7375)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Information Rate</td>
<td></td>
<td></td>
<td>0.3694</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spanish</th>
<th>observed</th>
<th>predicted</th>
<th>act-def</th>
<th>def</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>act-def</td>
<td></td>
<td>99</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>def</td>
<td></td>
<td>23</td>
<td>98</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>indefinite</td>
<td></td>
<td>14</td>
<td>28</td>
<td>74</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.715</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% CI</td>
<td></td>
<td></td>
<td>(0.6667, 0.76)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Information Rate</td>
<td></td>
<td></td>
<td>0.372</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Macedonian</th>
<th>observed</th>
<th>predicted</th>
<th>act-def</th>
<th>def</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>act-def</td>
<td></td>
<td>95</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>def</td>
<td></td>
<td>33</td>
<td>75</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>indef</td>
<td></td>
<td>12</td>
<td>43</td>
<td>71</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.6359</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% CI</td>
<td></td>
<td></td>
<td>(0.5852, 0.6844)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Information Rate</td>
<td></td>
<td></td>
<td>0.3694</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Russian</th>
<th>observed</th>
<th>predicted</th>
<th>act-def</th>
<th>def</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>act-def</td>
<td></td>
<td>92</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>def</td>
<td></td>
<td>30</td>
<td>87</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>indefinite</td>
<td></td>
<td>12</td>
<td>35</td>
<td>76</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.6728</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% CI</td>
<td></td>
<td></td>
<td>(0.6231, 0.7199)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Information Rate</td>
<td></td>
<td></td>
<td>0.3694</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
I showed how parallel movie subtitles can be used to examine the factors relevant to the expression of referentiality.

**Referential values**

- we can distinguish between 3 major values of referentiality taking into account the distribution of properties of the referential expressions

**Modelling referentiality**

- although in German, Spanish, and Macedonian, articles are by far the most important predictor for referentiality,
  - the indefinite article is more important than the definite article
  - other properties of the referential expression are important predictors as well (possessives, type of expression, demonstratives)
  - the models do not perform much worse if the article is taken out as a predictor
Thank you!
Examples

Even though the referential expressions are equivalent, they do not have to have the identical referential value in all languages:

(8)  
  a. And in the gloom of Gollum’s cave, it waited.
  b. Darkness crept back into the forests of the world.

sp  
  a. Y en la oscuridad de la cueva de Gollum, aguardó.
  and in the darkness of the cave of Gollum, waited.3sg
  b. La oscuridad se empezó a filtrar en los bosques del
     the darkness REFL started.3SG PREP infiltrate in the forests of the
     mundo. world

de  
  a. Und in der Finsternis von Gollums Höhle wartete er.
  and in the darkness of Gollum.GEN cave waited he
  b. Dunkelheit legte sich über den Wald der Welt.
     darkness laid REFL over the forest the.GEN world
Distribution of grammatical functions

grammatical functions across referential values

- Specific
- Nonspecific

Conclusion
Decision tree (Spanish)