Definiteness in languages with and without articles

Jingting Ye & Laura Becker

Fudan University & Leipzig University

13th Conference on Typology and Grammar for Young Scholars
24.11.2016
ILS RAN, St Petersburg
Introduction

(In)definiteness and specificity

Outline: A pilot study based on multiple parallel movie subtitles

Results

- Comparing the use of articles
  - Examples
  - Factor strength based on random forests
- Other strategies to mark (in)definiteness
  - Demonstratives
  - Adnominal indefinites
  - The numeral one
  - Word order
- The levels of givenness
- Relevant factors for definiteness
  - Trees
  - Random forests

Concluding remarks
There are many accounts for definiteness, however, most rely on language-specific expressions, e.g. definite articles.

Although comparative studies exist, no empirically based cross-linguistic study seems to be available yet that makes expressions of definiteness comparable directly.

This pilot study explores the possibilities of parallel texts for comparing the expression of definiteness in languages with and without articles.

The languages we examined are German, Hungarian, Russian, and Chinese.
Definiteness has been associated with the following concepts:

- uniqueness (Frege 1892; Strawson 1950; Heim & Kratzer 1998; Stanley & Gendler Szabó 2000)
- familiarity (Heim 1988; Roberts 2003; Chierchia 1995)
- identifiability (e.g. Birner & Ward 1994; Schroeder 2011)
- anaphora (e.g. Ariel 1988, 2001) and bridging (Clark 1975)
- quantification (Löbner 1985; Kamp 2002)
Definiteness (Dryer 2013, 2014)

(main focus: classification of articles)

Reference hierarchy

anaphoric definites  definite noun phrases that refer back in the discourse
non-anaphoric definites  based on shared knowledge of the speaker and hearer
pragmatically specific indefinites  subsequent reference, introduce a participant into the discourse that is referred to again in the subsequent discourse
semantically specific  associated with an entailment of existence
semantically nonspecific  not associated with an entailment of existence
Outline of the present study
Outline of the study

- Comparison of the coding strategies for (in)definiteness in four languages based on parallel texts.
- Parallel texts are necessary, as they ensure that the situations of use are directly comparable in the different languages.
- The corpus: 4 movies (Inception, Noah, Frozen, Avatar)
- From those subtitles, 295 referring expressions with sufficient similarity have been extracted for German, Hungarian, Chinese, and Russian.
- In total, we annotated 1180 tokens.
Annotation

- **noun.type** count, mass, proper, person, place
- **animacy** y, n
- **givenness** referential properties (definiteness)
  - def.d: deictic
  - def.a: anaphoric
  - def.su: situational unique
  - spec.p: pragmatically specific
  - spec.s: semantically specific
  - non.spec: non-specific
  - generic
- **synt.pos** S, O, obl, attr, pred
- **article** def, indef, n
- **poss** y, n
- **dem** y, n
- **adj** y, n
- **other.attr** y, n
- **bare.noun** y, n
- **pronoun** n, y, drop
- **number** sg, pl

y=yes, present
n=no, not present
The use of articles
Article frequencies
Bare noun vs. indefinite article

In predicative uses, Hungarian often lacks the indefinite article.

(1)  
I am a man.

de  Ich bin ein Mann.

hu  Ember vagyok.

zh  wo shi yi ge ren.

ru  Я человек!
Definite article vs. adnominal demonstrative

German uses the definite article in contexts, where the other languages require a demonstrative.

(2) That allows us to get right in the middle of that process.

De: Das erlaubt uns, mitten in den Prozess

That allows us in the middle into the process einzusteigen.

Enter

Hu: ez az, amiért bele tudunk szólni ebben a folyamatban.

This that why in can.2PL say this in the process.

Enter

Zh: women jiu neng zhijie jinru zhe ge guocheng.

We ADV can directly get into this process.

Ru: Это позволяет нам проникать внутрь этого

This allows us permeate inside. ACC this GEN процесса.

Process. GEN
Factors determining the use of articles

Random forests (e.g. Baayen & Tagliamonte 2012; Baayen et al. 2008) can help to determine the strength of factors, i.e. how much those properties are correlated with the uses of articles.

What are random forests?

- Random forests are based on a large number of conditional inference trees of random sub-samples of the data.
- Trees split the data according to the factor that makes the purest groups with the smallest p value with respect to the value that we want to test (article).
- Growing a large number of trees allows to control for factors that depend on each other and smaller effects, otherwise hidden by more influential factors, can also be considered.

The following factors have been considered here:

```plaintext
art ~ synt.pos + poss + dem + adj + other.attr + number + noun.type + animacy + givenness
```
Factors determining the use of articles

Based on the random forest model, conditional variable importance measures the importance of the factors:

The values for a factor are randomly manipulated. The greater the impact on the accuracy of the prediction, the more important the factor is. If the random values of a factor do not influence the accuracy of the prediction of the model, the factor has no significant overall influence (to the left of the red line).
Factors determining the use of articles

- In both languages, definiteness determines most the use of articles.
- Also, the semantic properties of the referent (noun.type) play a role.
- Possessive markers are relevant in German, since they formally require the absence of an article.
- In German, number is relevant as well; in Hungarian, on the other hand, number does not seem to have a great impact on the use of the article. (This difference cannot directly be explained by observed frequencies of certain combinations and would need to be addressed in more detail.)
- Other factors have less or no significant influence on the presence/choice of the article.
Factors determining the use of articles

How well do the random forest models predict the data?

<table>
<thead>
<tr>
<th>Confusion matrix <strong>German</strong></th>
<th>Confusion matrix <strong>Hungarian</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
</tr>
<tr>
<td>Prediction</td>
<td></td>
</tr>
<tr>
<td>def</td>
<td>70</td>
</tr>
<tr>
<td>indef</td>
<td>12</td>
</tr>
<tr>
<td>n</td>
<td>19</td>
</tr>
</tbody>
</table>

Accuracy: 0.7966
No Information Rate: 0.4712

Accuracy: 0.7119
No Information Rate: 0.4712

Why look at those predictions?

- Results about the importance of factors are based on the models used.
  
- Therefore, it is crucial to take into account to what extent the model used actually captures the data as it is observed.

- In this case, the models capture around 70% (Hungarian) to 80% (German) of the data, being clearly above the no information rate (the proportion of what is correctly predicted by allocating the observations to the three values randomly).
Marking givenness without articles
Demonstratives

Chinese and Russian can use a demonstrative in immediate anaphoric (but not deictic) contexts, where German, Hungarian, and English feature a definite article.

(3) Who are the people?
de Wer sind die Leute?
who are the people
hu És kik az emberek?
and who the people
zh zhe xie ren shi shei?
this CL people COP who
ru Кто эти люди?
who these people
Demonstratrices

Chinese and Russian can use a demonstrative in immediate anaphoric (but not deictic) contexts, where German, Hungarian, and English feature a definite article.

(4) Well, down in the dream, Mal showing up.

de Na ja, weil im Traum Mal aufgetaucht ist. well because in the dream Mal show up. PTCP is

hu Csak mert álimodban Mal megint jelen volt. only because dream poss:2sg in Mal again present was

ru Просто в том сне появилась Мол. simply in that dream showed up Mal
Adnominal indefinites (Russian)

(from a different dataset)
In contexts where other languages (en, de, da, fr, sp, ro, hu) use an indefinite article, Russian marks the referent as not yet identifiable by an adnominal indefinite. Other languages without indefinite articles (mk, bg, ee, fi) simply feature a bare noun.

(5)  A boy! There’s a boy in the water!

de  Seht mal. Ein Junge ist im Wasser..
da  Der er en dreng i vandet!
fr  Un garçon! Un garçon, sur l’eau!
sp  ¡Un niño! ¡Hay un niño en el agua!
ro  Un băiețel. E un băiețel în apă.
hu  Egy fiú van a vízben!
mk  Погледнете, момче! Момче во водата!
bg  Вижте, момче във водата!
ru  Там, в воде! Какой-то мальчик!
ee  Vaata, poiss! Seal vees on poiss!
fi  Vedessä on poika!
The numeral **one** (Russian)

(from a different dataset)

Russian, amongst other languages without indefinite articles (mk, bg, ee, fi), use the numeral *one* to introduce a pragmatically specific referent.

(6) *I just heard about this great place.*

de *Ich weiß einen ganz tollen Ort.*
da *Jeg har lige hørt om et skønt sted.*
fr *J’ai entendu parler d’un super endroit.*
sp *Ven. Sé de un lugar fantástico.*
ro *Am auzit despre un loc grozav.*
hu *Tudok egy állati jó helyet.*
mk *Слушнав за едно много добро место.*
bg *Току-що чух за едно страхотно място.*
ru *Пошли! Мне тут про одно место рассказали.*
ee *Tule, ma kuulsin ühest põnevast kohast.*
fi *Tule. Kuulin yhdestä hienosta paikasta.*
Chinese uses the numeral one much more frequently than Russian to express non-identifiability or non-specificity. In most instances, the referent is in object or predicate position.

The numeral one marking non-specificity:

(7) Imagine you’re designing a building.

de Sie entwerfen ein Gebäude.
you design a building

hu Tegyük fel, tervezel egy házat.
let’s assume plan.2SG a house.ACC

zh ni sheji yi zuo jianzhuwu shi
you plan one CL building if

ru Представь, что ты проектируешь здание.
imagine.IMP COMPL you plan building
The numeral *one* (Chinese)

In this example, the numeral *one* has been omitted; however, the general classifier *ge* is still present marking generic use:

(8) You ’re a gift.

*de* Du bist ein Geschenk.
you are a gift

*hu* Te egy ajándék vagy.
you a gift are

*zh* ni shi ge liwu
you COP CL gift

*ru* Потому что ты дар.
because you gift
Word order

Discourse-old, i.e. given referents tend to occur sentence-initially.

(9) **The storm** cannot be stopped.

de **Der Sturm** kann nicht aufgehalten werden.

hu **A vihart nem lehet megállítani.**

zh **baofengyu shi wufa zuzhi de.**

ru **Бурю нельзя остановить.**
Word order

Discourse-new, i.e. not given referents tend to occur sentence-finally.

(10) That’s where they should be. They have a purpose.

```
de Sie haben einen Zweck.
they have a purpose
```

```
hu Céljuk van.
goal.POSS:3PL is
```

```
zh tamen zhi you yi ge mudi
they only have one CL purpose
```

```
ru У них есть цель.
at them is purpose
```
Word order

In this examples, a pragmatically specific referent is introduced. English, German, and Hungarian use an indefinite article, but keep the referent in subject position and sentence initially.

In both Chinese and Russian, no marker is used. However, the referent appears sentence finally, in Russian also as subject.

(11) **A great flood** is coming.

de **Eine große Flut** kommt.
a great flood comes

hu **Hatalmas vízözön** közeleg.
great flood approaches

zh **daoshi hui you** **hongshui**
then fut have flood

ru **Близится Великий потоп.**
approaches great flood
The levels of givenness
Counts of the givenness levels

![Bar chart showing counts of givenness levels](chart.png)
Clustering the givenness levels

**definite**  
def.d, def.a, def.su

**indefinite**  
spec.p, spec.s, generic, nonspec

The distance of the levels is measured based on:
synt.pos, article, possessive, classifier, demonstrative, adjective, other attribute, pronoun, number
Clustering the givenness levels

definite  def.d, def.a, def.su
indefinite  spec.p, spec.s, generic, nonspec

The distance of the statuses is measured based on:
synt.pos, article, possessive, classifier, demonstrative, adjective, other attribute, pronoun, number
The levels of givenness

**Clustering the givenness levels**

**definite**  
def.d, def.a, def.su

**indefinite**  
spec.p, spec.s, generic, nonspec

The distance of the statuses is measured based on:
synt.pos, article, possessive, classifier, demonstrative, adjective, other attribute, pronoun, number
Clustering the givenness levels

- Considering all 4 languages, definites and indefinites fall into two major clusters.
- This holds for the individual languages as well, except in Russian, where non-specific clusters with anaphoric and deictic definites, as well as situational unique definites with specific and generic.
- Situational unique definites consistently pattern less with anaphoric and deictic definites.
- Instead two levels, three levels of definiteness could be distinguished:
  - **indef**, **def1** (anaphoric, deictic), **def2** (situational unique)
- The cluster of indefinites is less consistent. A larger data set might help to say more about that.
Relevant factors for definiteness
Relevant factors

A note on the levels of definiteness

- As clustering has shown, the data from all 4 languages provides evidence for a split of the definite values into anaphoric and deictic on the one hand (def1), and situational unique on the other (def2).
- This distinction of three values for definiteness (def1, def2, indef) has been maintained for the analysis of relevant factors based on trees and forests.
- Having only two values for the variable tested gives a higher accuracy (the model only has to be able to predict one of two possible outcomes) than with three values.

→ Why not a two-way distinction of def vs. indef?
→ Although the model predicts better with only two values for definiteness (higher accuracy), it cannot capture the impact of factors that are relevant to distinguish between def1 and def2. Considering these as separate values gives a more accurate picture of what is linguistically relevant to express definiteness.
Relevant factors: conditional inference trees

German

Results

Relevant factors for definiteness

Jingting Ye & Laura Becker

Definiteness and articles

35 / 49
Relevant factors: conditional inference trees

Hungarian
Relevant factors: conditional inference trees

Chinese
Relevant factors for definiteness

Relevant factors: conditional inference trees

Russian

Node 2 (n = 40)

Node 5 (n = 18)

Node 6 (n = 167)

Node 8 (n = 19)

Node 9 (n = 51)
Relevant factors: conditional inference trees

NB: Conditional inference trees provide rather a first approximation and a visualization of the interaction and strength of the factors tested for definiteness (for more reliable results, we will consider forests next).

German and Hungarian
- The article is the most influential factor for definiteness.
- The semantic properties of the noun are not very influential; only for a sub-part of the data, proper nouns and places are correlated with the definite and all other values with the indefinite values.
- The syntactic position is not relevant.

Chinese and Russian
- The most relevant factors are possessive markers and the semantic properties.
- The syntactic position of the referent is relevant to definiteness for persons, proper nouns, and place names.

In all languages
- Demonstratives and pronouns are relevant for def1, while possessives often mark def2.
Relevant factors: random forests

Strength of the factors (conditional variable importance)
Relevant factors: random forests

How well do the models predict the data?

Confusion matrix **German**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>def1</th>
<th>def2</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>def1</td>
<td>50</td>
<td>21</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>def2</td>
<td>22</td>
<td>58</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>indef</td>
<td>15</td>
<td>3</td>
<td>105</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: 0.722
No Information Rate: 0.4271

compared to:
Accuracy with 2 levels of definiteness: 0.8949
Accuracy with 6 levels of definiteness: 0.4983

Confusion matrix **Hungarian**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>def1</th>
<th>def2</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>def1</td>
<td>60</td>
<td>22</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>def2</td>
<td>12</td>
<td>46</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>indef</td>
<td>15</td>
<td>14</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: 0.6746
No Information Rate: 0.4271

compared to:
Accuracy with 2 levels of definiteness: 0.8034
Accuracy with 6 levels of definiteness: 0.4373

- In both cases, accuracy is about 70%.
- In German, we find confusion mainly between def1 and def2, but below a random level.
Relevant factors: random forests

Strength of the factors (conditional variable importance)
## Relevant factors: random forests

How well do the models predict the data?

### Confusion matrix Chinese

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>def1</th>
<th>def2</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>def1</td>
<td>37</td>
<td>11</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>def2</td>
<td>9</td>
<td>35</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>indef</td>
<td>41</td>
<td>36</td>
<td>112</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: 0.6237  
No Information Rate: 0.4271

compared to:

Accuracy with 2 levels of definiteness: 0.7898  
Accuracy with 6 levels of definiteness: 0.4678

### Confusion matrix Russian

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>def1</th>
<th>def2</th>
<th>indef</th>
</tr>
</thead>
<tbody>
<tr>
<td>def1</td>
<td>42</td>
<td>7</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>def2</td>
<td>8</td>
<td>40</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>indef</td>
<td>37</td>
<td>35</td>
<td>114</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: 0.6644  
No Information Rate: 0.4271

compared to:

Accuracy with 2 levels of definiteness: 0.7356  
Accuracy with 6 levels of definiteness: 0.4678

- In both cases, accuracy is above 60% and the no information rate.
- In contrast to what we found in German and Hungarian, in both Chinese and Russian, *indefinites* are confused with both *def1* and *def2*, while there is less confusion between the latter two values (at least in Russian)
Relevant factors: random forests

Summing up the strength of the factors

- Random forests largely support the impact of the factors seen in trees:
- The article is most relevant in both German and Hungarian.
- In Chinese and Russian, the most important clue wrt to definiteness comes from possessive marking, semantics, and demonstratives.
- In the two languages with articles, demonstratives play a less significant role.
- The syntactic position is not relevant in the two languages with articles, being significant in both Russian, and to a greater extent, in Chinese.
- Pronouns are relevant in all languages but Chinese (why this is the case needs further investigation).
- Also the presence of a classifier (as seen, e.g., in combination with a numeral) shows significant impact on the marking of definiteness in Chinese.
- Factors that do not show any influence on definiteness are: presence of an adjective, other attributes (except for Chinese).
- Number seems to have a statistically relevant impact in German and Chinese, but not in Russian and Hungarian.
This pilot study explored parallel texts for comparing the expression of givenness/definiteness across languages, including languages without articles.

+art In both German and Hungarian, as expected, the article is the most important factor wrt the expression of givenness. The interaction and importance of other factors, however, differs, as well as the factors determining the use of articles. Using parallel texts, those two kinds of variation are made directly comparable.

-art Although both Russian and Chinese have no articles, we saw differences in coding strategies for values of givenness, e.g. demonstratives, and the numeral one, a potentially emerging indefinite article in Chinese. While bare nouns tend to be interpreted as indefinite/non-specific in languages with articles, they seem to be rather definite by default in Russian, so that strategies of “downgrading” (e.g. adnominal indefinites) are used in contexts of non-identifiability.
Concluding remarks II

±art While some properties set apart German and Hungarian from Chinese and Russian (e.g. the influence of demonstratives, syntactic position), other properties grouped German and Chinese (influence of number) or set apart Chinese from the rest (influence of pronominal use). This suggests that languages do not strictly fall into two clusters depending on the presence/absence of articles with respect to their encoding of givenness. To what extent languages pattern together into those two or other groups seems to be worth pursuing in further research using the approach presented here.

Levels of definiteness Clustering the levels of givenness according to their encoding in the four languages revealed two major clusters (def, indef) in both languages with and without articles. Also, all languages showed a difference between anaphoric/deictic and non-anaphoric/deictic definites, motivating a three-way distinction of levels of definiteness. Including more languages will yield a more fine-grained picture of cross-linguistically relevant categories of givenness.
References I

References II


Thank you!