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

- 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

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

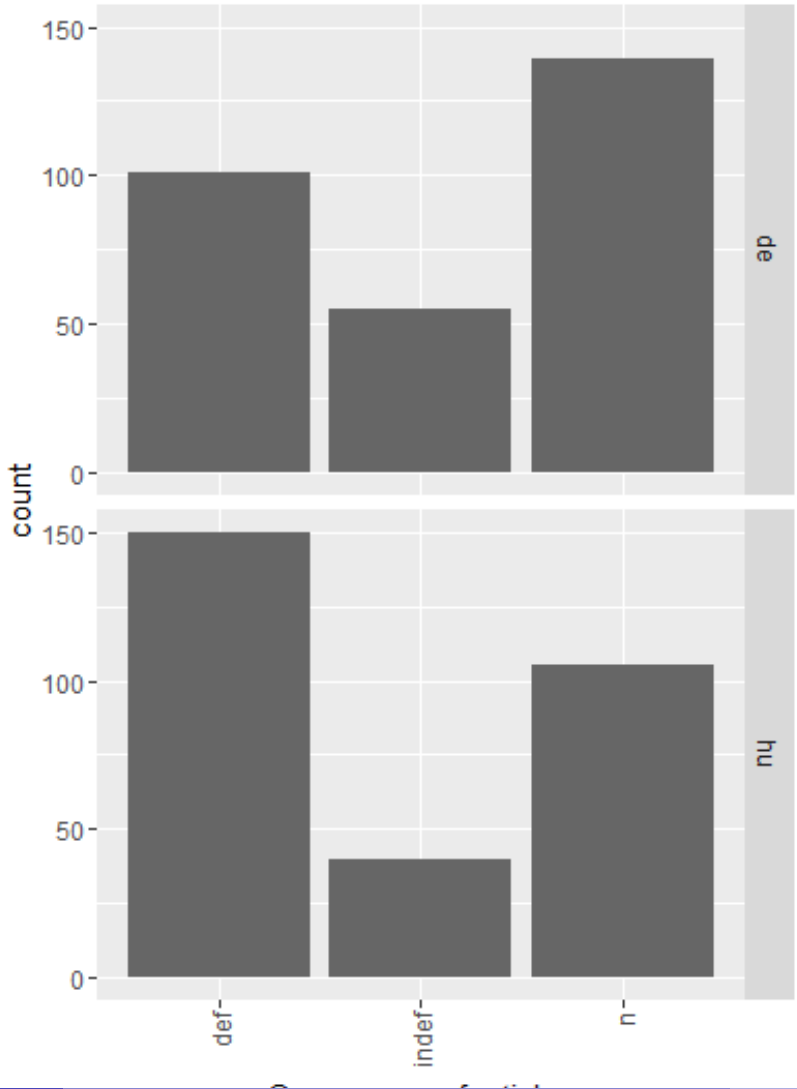
- Comparison of the coding strategies for (in)definiteness in four languages based on parallel texts.
- Parallel texts are necessary, as they ensure that the situations of use are directly comparable in the different languages.
- The corpus: 4 movies (Inception, Noah, Frozen, Avatar)
- From those subtitles, 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

The use of articles

Article frequencies



Bare noun vs. indefinite article

predicative use

(1) *I am* ***a*** ***man***.

de *Ich bin* ***ein*** ***Mann***.

hu ***Ember*** *vagyok*.
man am

zh *wo shi* ***yi*** ***ge ren***.
I COP one CL man

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*** *einzusteigen.*

hu *ez az, amiért bele tudunk szólni* ***ebbe a*** ***folyamatba***.
this that why in can.2PL say this.in the process.in

zh *women jiu neng zhijie jinru* ***zhe ge guocheng***.
we ADV can directly get.into this CL process

ru *Это позволяет нам проникать внутрь* ***этого*** ***процесса***.

Factors determining the use of articles

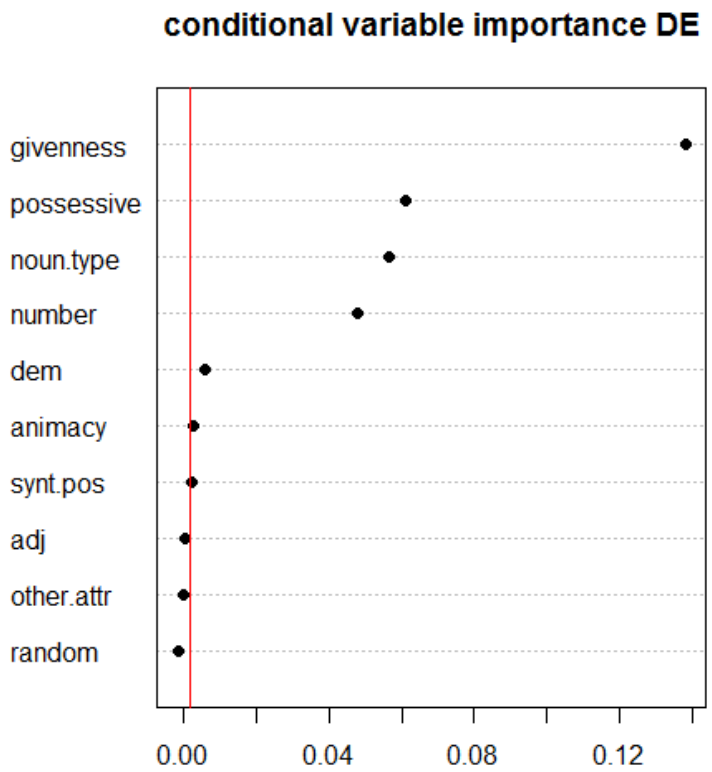
Random forests (e.g. Baayen & Tagliamonte 2012; Baayen et al. 2008) can help to determine the strength of factors, i.e. how much those properties are correlated with the uses of articles.

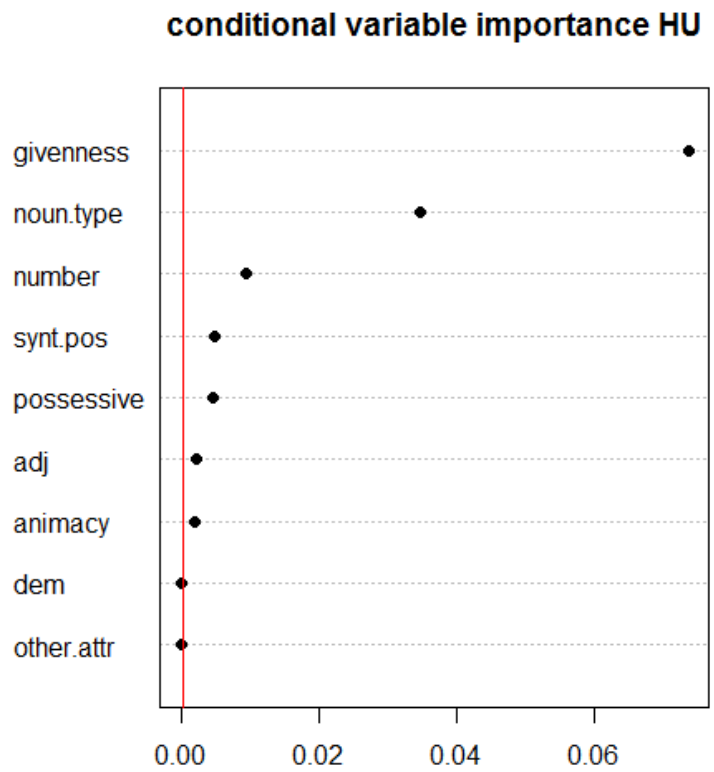
What are random forests?

- Random forests are based on a large number of conditional inference trees of random subsamples 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 influencial factors can also be considered.

Factors determining the use of articles

art ~ synt.pos + possessive + dem + adj + other.attr + number + noun.type + animacy + givenness





Accuracy : 0.8

Accuracy : 0.7051

Marking givenness without articles

Demonstratives

anaphoric use

(3)
*Who are **the** **people**?*

de
*Wer sind **die** **Leute**?*

hu
*És kik **az** **emberek**?*
and who the people

zh
***zhe** **xie** **ren** shi shei?*
this CL people COP who

ru
*Кто **эти** **люди**?*

Demonstratives

anaphoric use

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

de *Na ja, weil **im** **Traum** Mal aufgetaucht ist.*

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

ru *Просто в **том** **сне** появилась Мол.*

Adnominal indefinites (Russian)

(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**!*

(Example from a different dataset)

The numeral *one* (Russian)

pragmatically specific referents

(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 многу 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****.*

(Example from a different dataset)

The numeral *one* (Chinese)

indefinite (non-specific) use in object position

(7) *Imagine you're designing **a** **building**.*

de *Sie entwerfen **ein** **Gebäude**.*

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**.*

The numeral *one* (Chinese)

in predicative use

(8)
- *You 're **a** **gift**.*
- de *Du bist **ein** **Geschenk**.*
- hu *Te **egy** **ajándék** vagy.*
  you a gift are
- zh *ni shi **ge** **liwu***
  you COP CL gift
- ru *Потому что ты **дар**.*

Word order

Discourse-old elements tend to occur sentence-initially.

(9)
de **Der Sturm** *kann nicht aufgehalten werden.*
hu **A vihart** *nem lehet megállítani*.
the storm.ACC NEG possible stop
zh ***baofengyu*** *shi wufa zuzhi de*.
storm COP cannot stop
ru ***Бурю*** *нельзя остановить*.

Word order

Discourse-new elements tend to occur sentence-finally.

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

de *Sie haben **einen** **Zweck**.*

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**.*

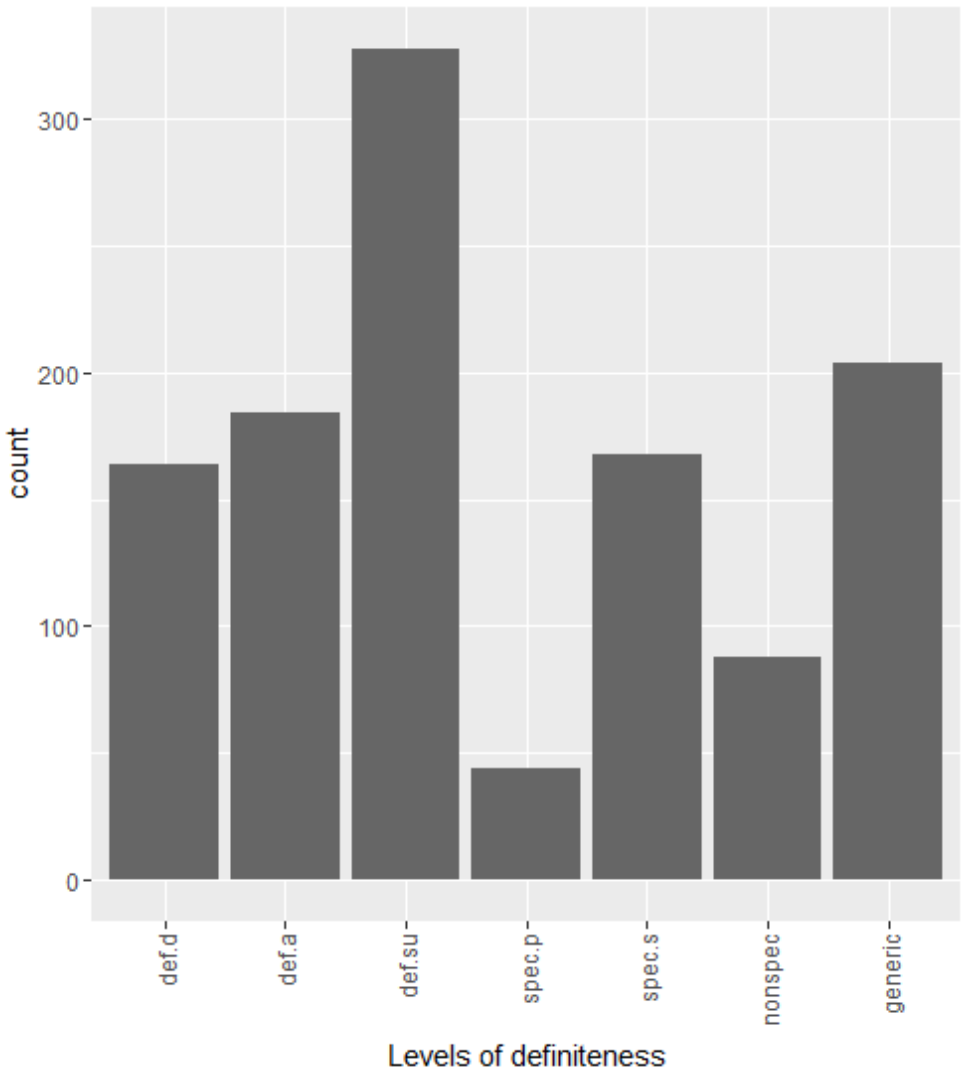
Word order

pragmatically specific

(11)
- ***A great flood*** *is coming.*
- de ***Eine große Flut*** *kommt.*
- hu ***Hatalmas vízözön*** *közeleg*.
  great flood approaches
- zh *daoshi hui you* ***hongshui***
  then fut have flood
- ru *Близится* ***Великий потоп***.

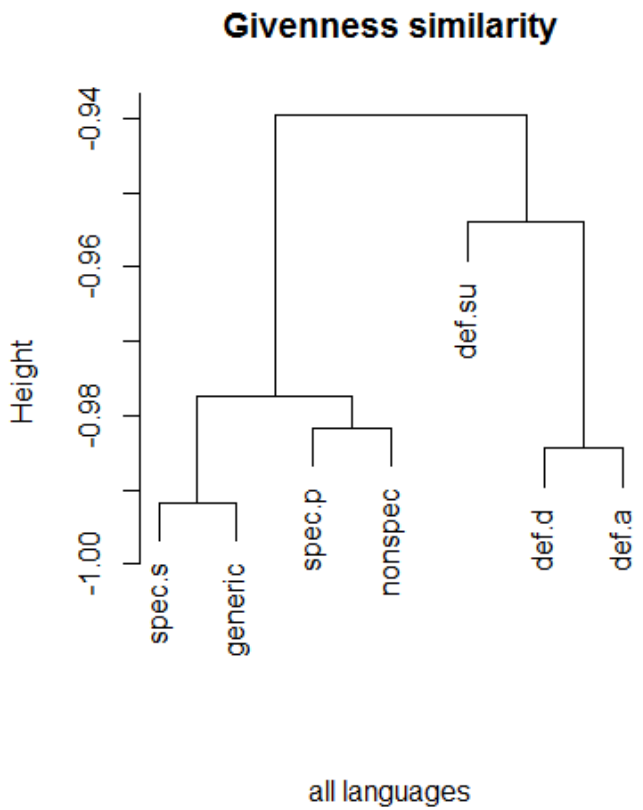
The levels of givenness

Counts of the givenness levels



Clustering the givenness levels

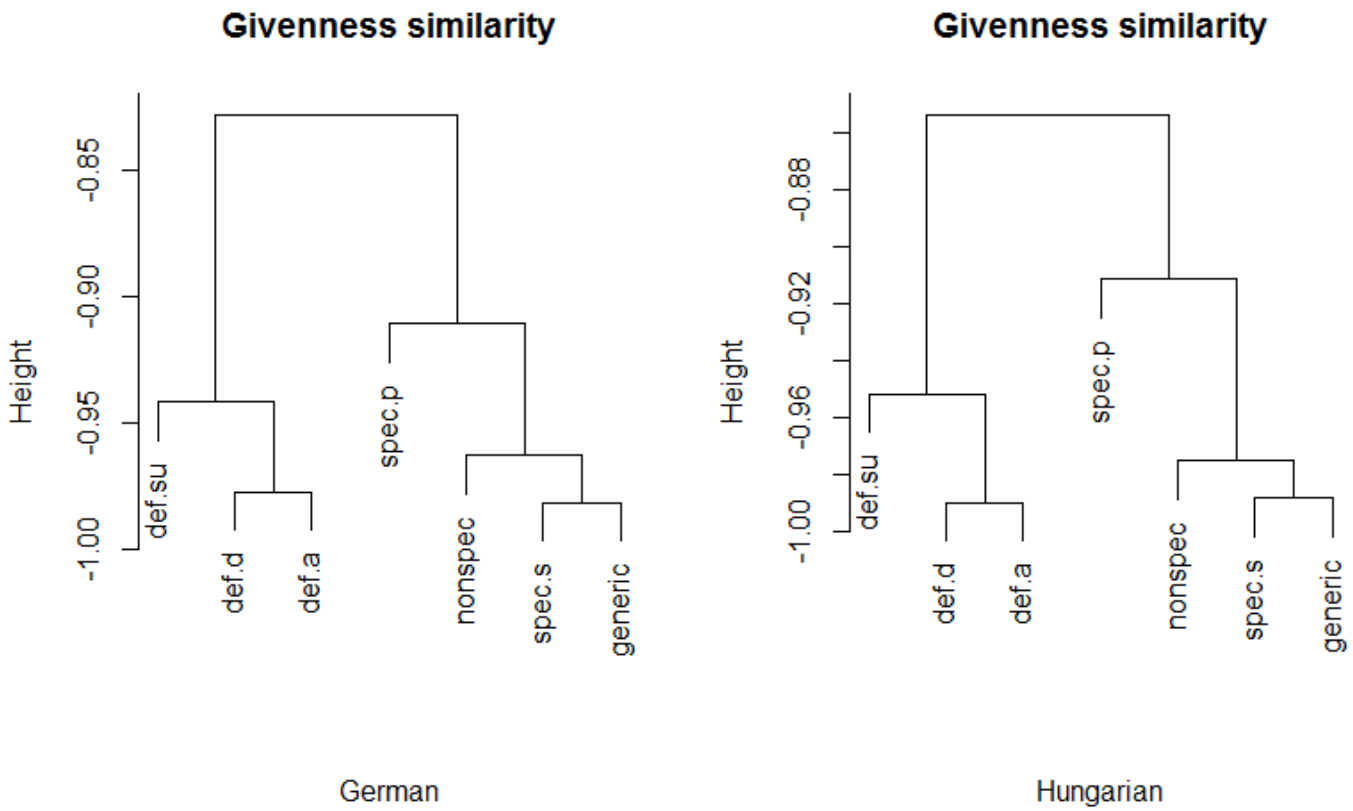
definite def.d, def.a, def.su
indefinite spec.p, spec.s, generic, nonspec



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

Clustering the givenness levels

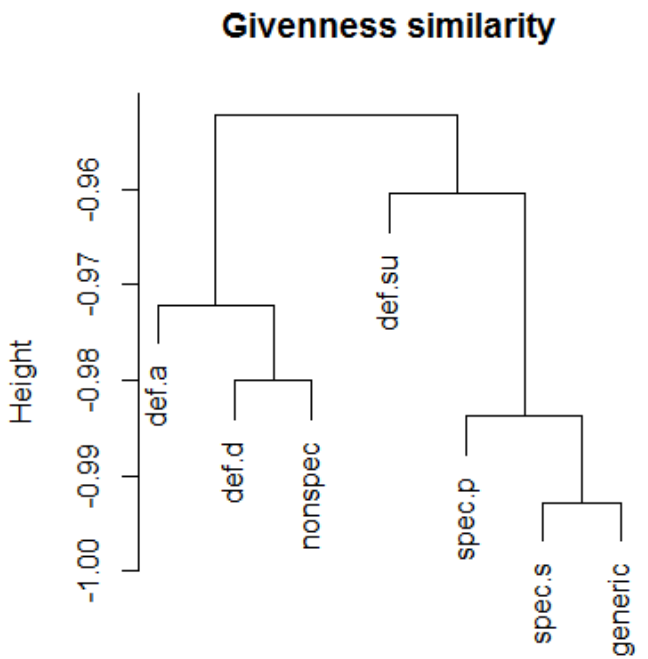
definite def.d, def.a, def.su
indefinite spec.p, spec.s, generic, nonspec



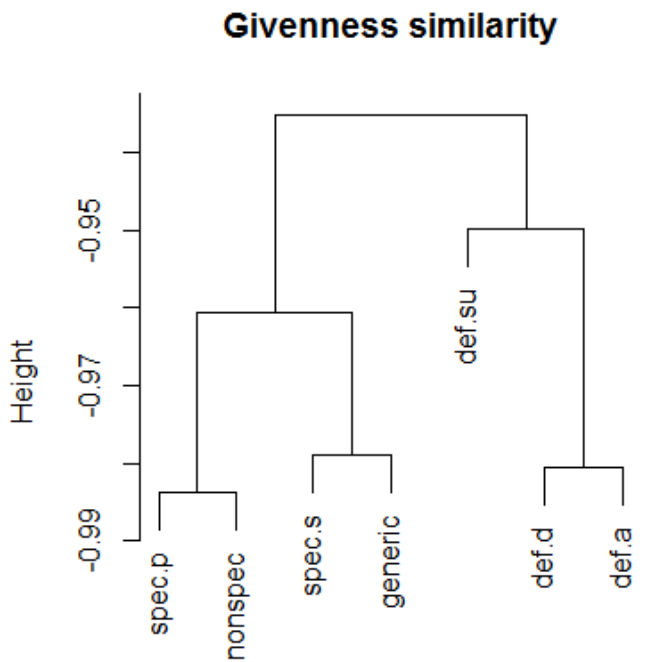
The distance of the statuses is measured based on:
synt.pos, article, possessive, classifier, demonstrative, adjective, other attribute, pronoun, number

Clustering the givenness levels

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



Russian

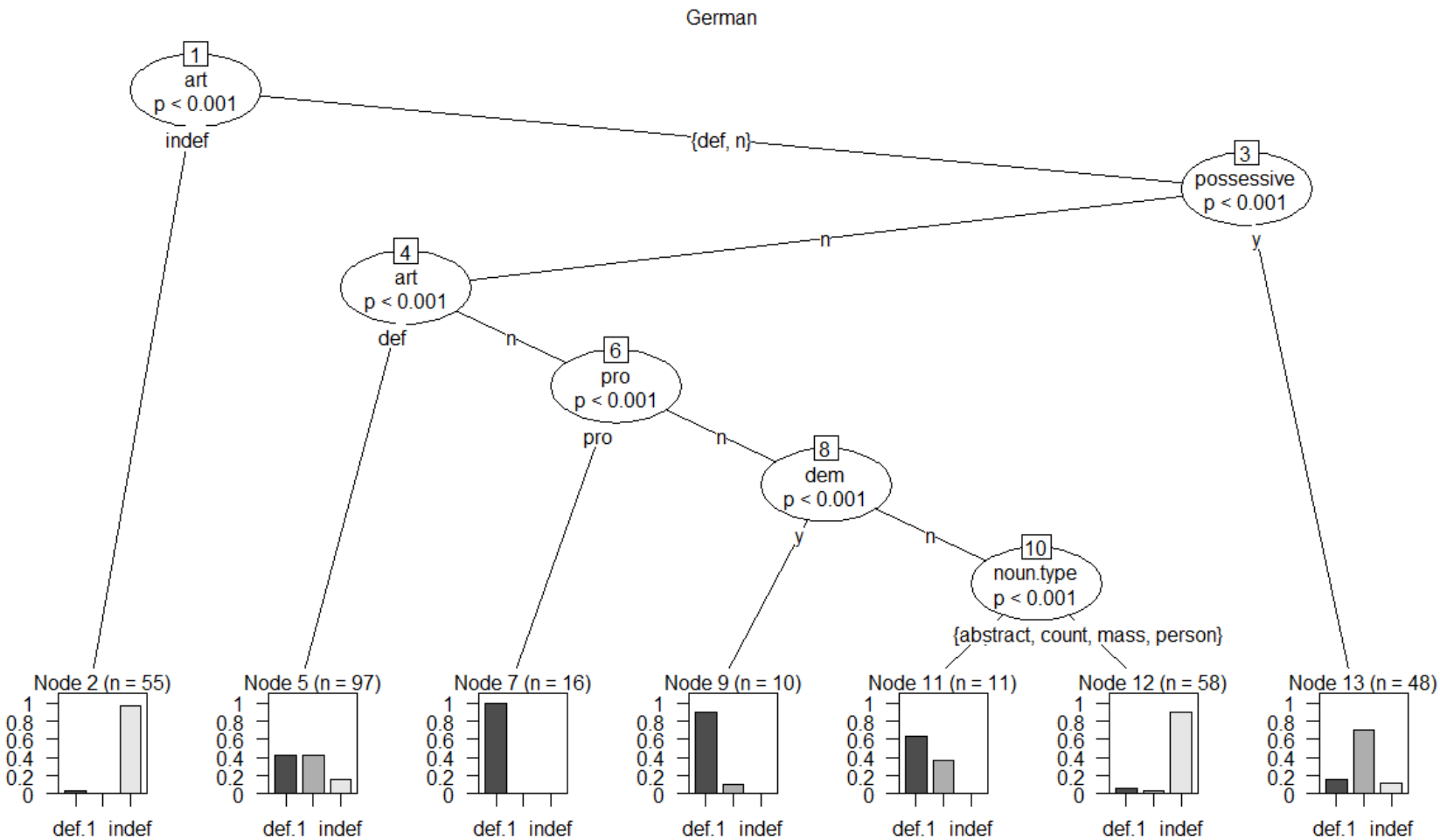


Chinese

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

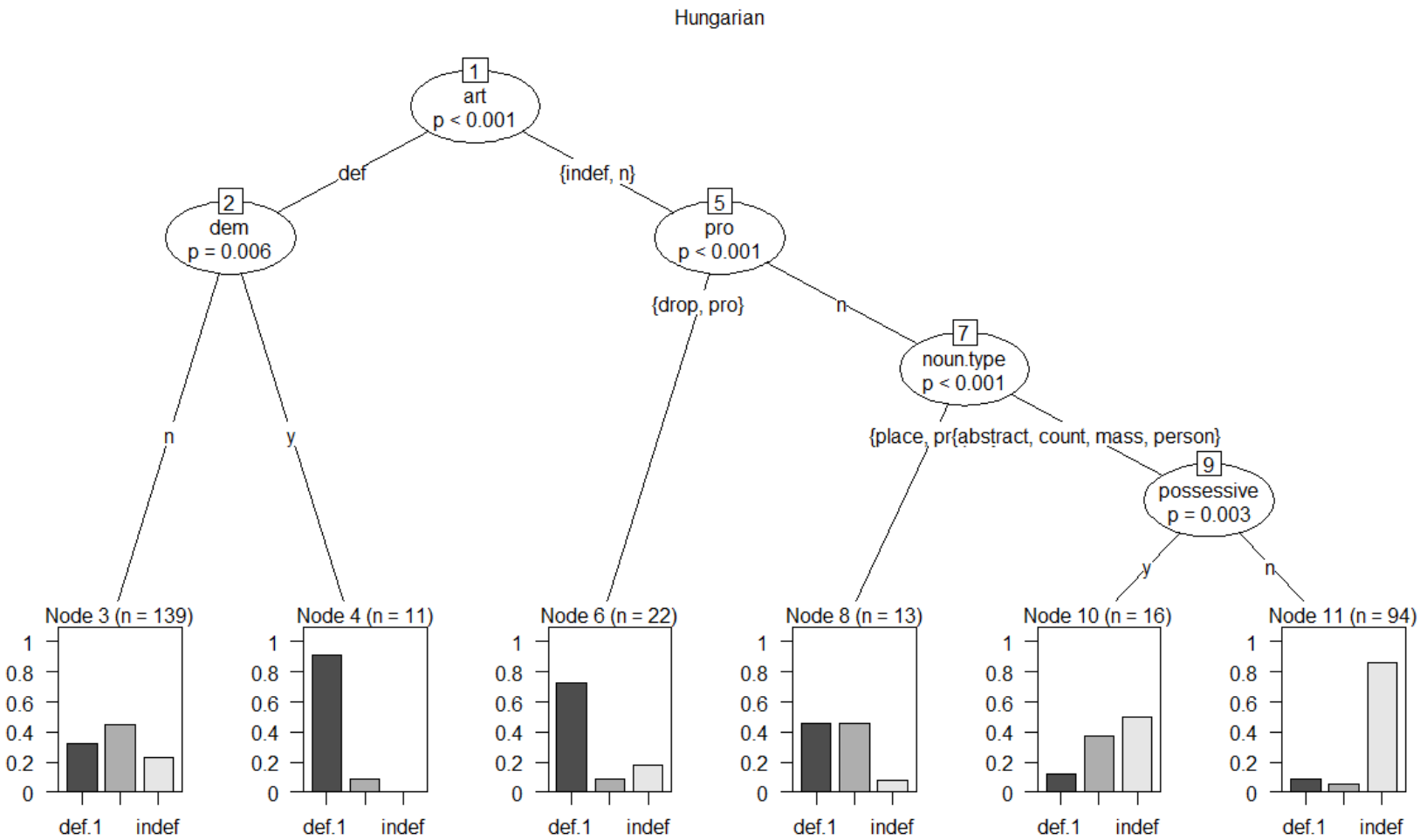
Relevant factors in German

3 Levels of givenness: def1, def2, indef (specific, non-specific)



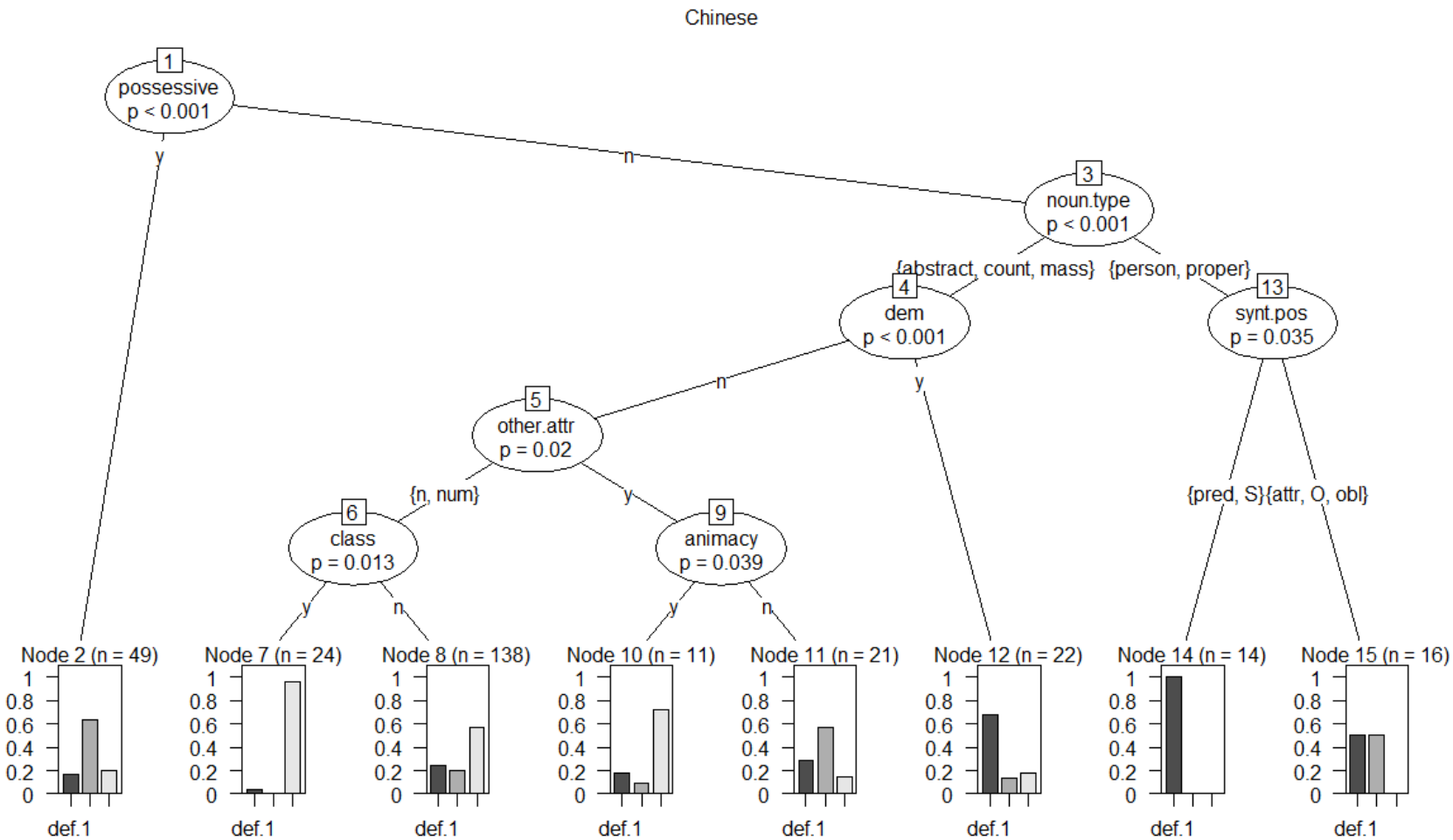
Relevant factors in Hungarian

3 Levels of givenness: def1, def2, indef (specific, non-specific)



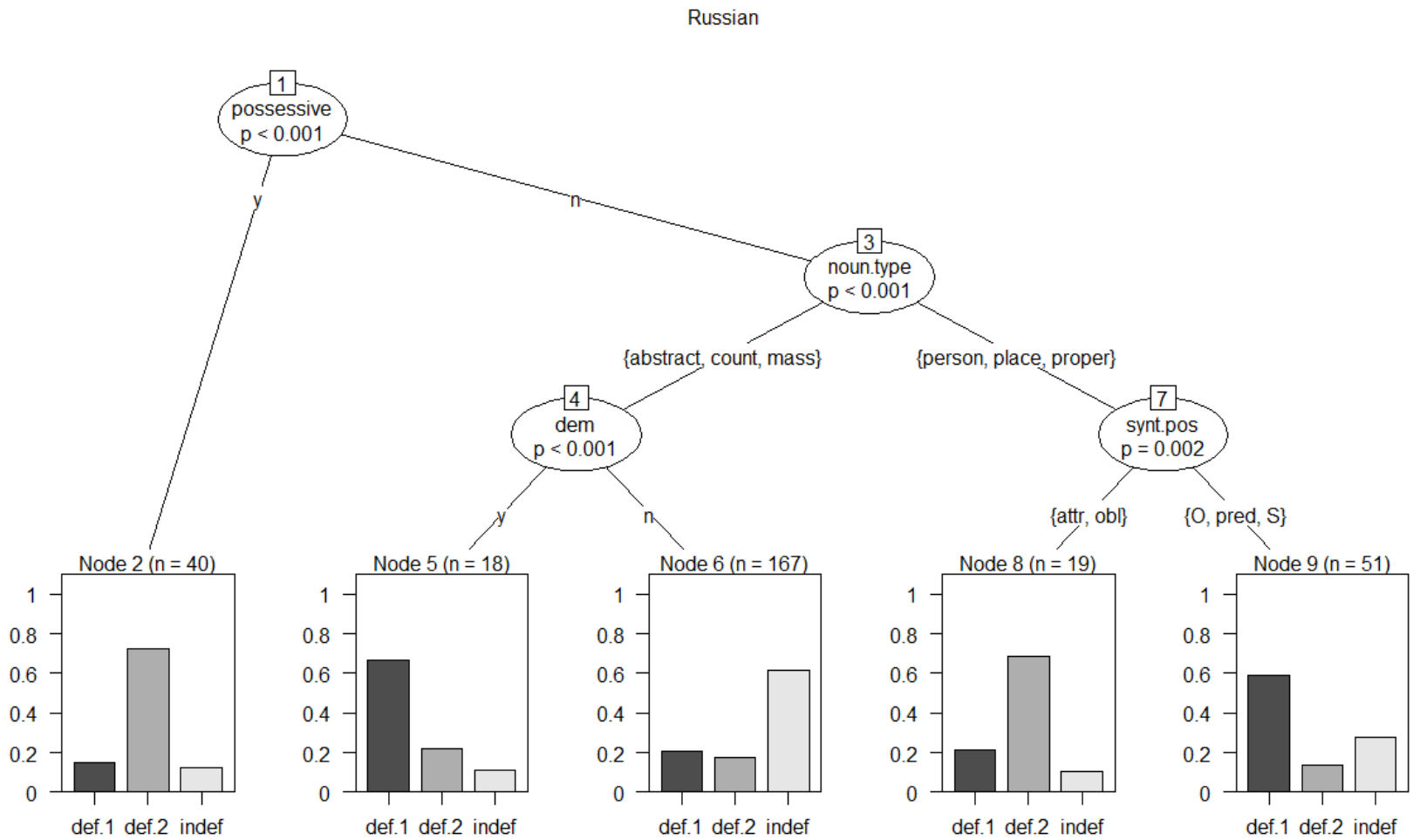
Relevant factors in Chinese

3 Levels of givenness: def1, def2, indef (specific, non-specific)

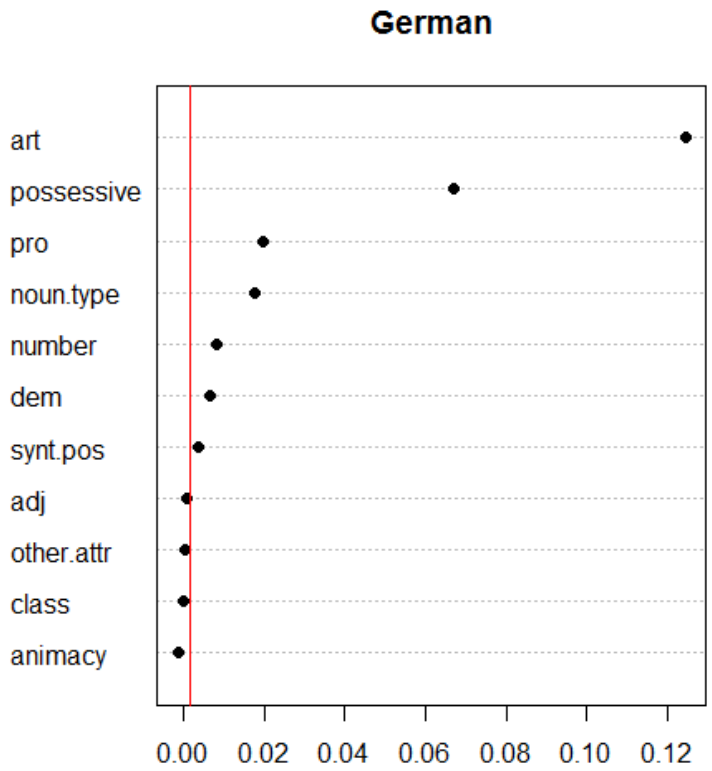


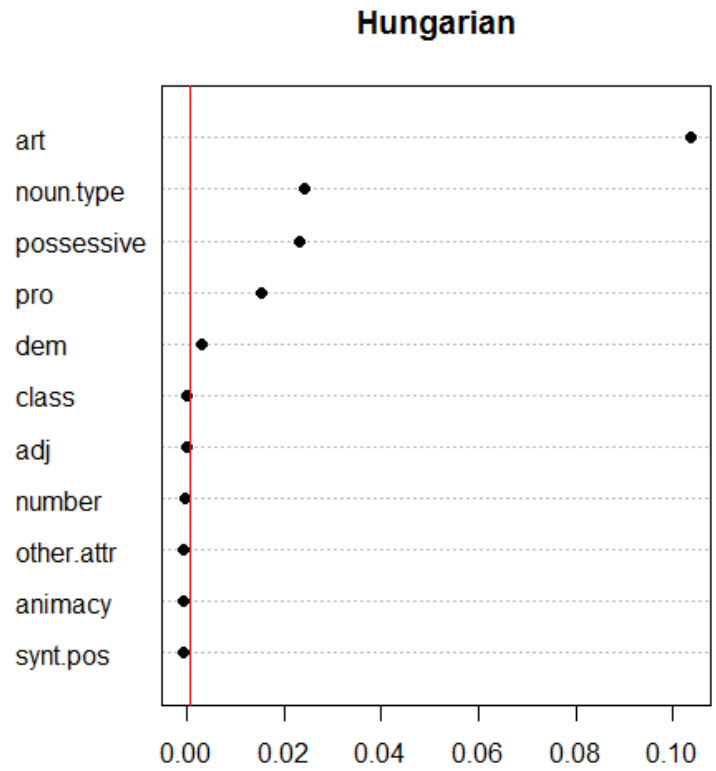
Relevant factors in Russian

3 Levels of givenness: def1, def2, indef (specific, non-specific)



Random forests

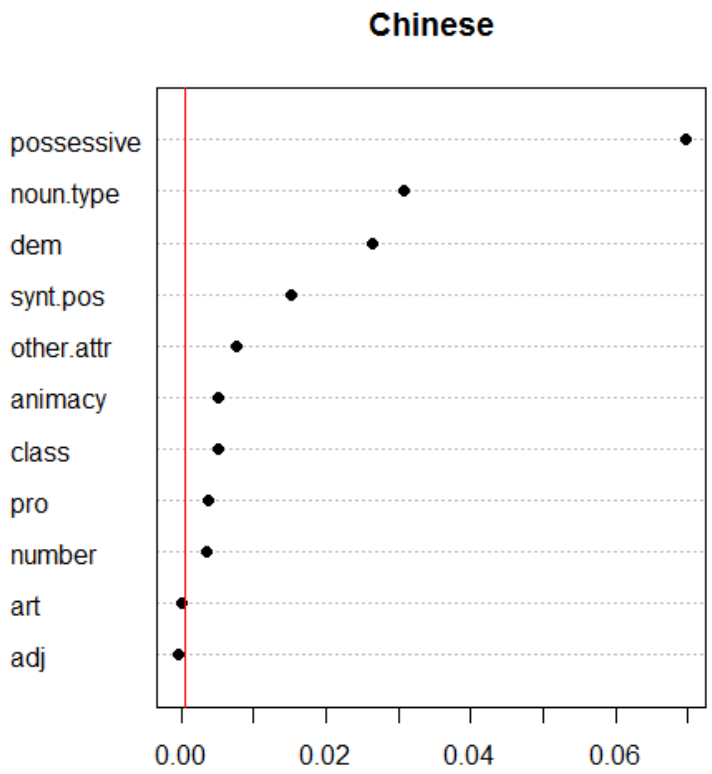


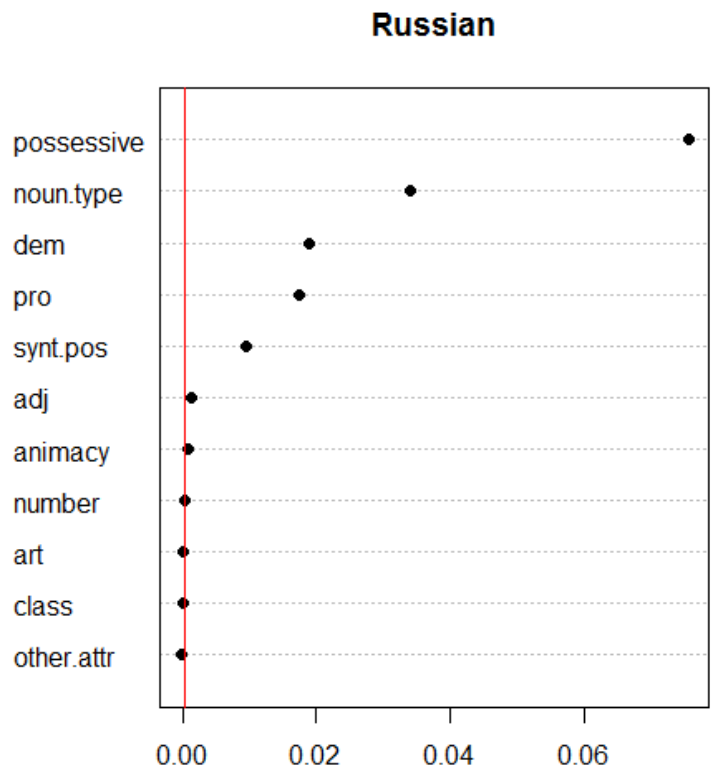


Accuracy :0.7254

Accuracy : 0.6712

Random forests





Accuracy : 0.6339

Accuracy : 0.6678

To sum up

- This pilot study explored parallel texts for comparing the expression of definiteness across languages, including languages without articles as primary means to encode reference status.
- **Articles** Referential properties have the highest impact on the use of the article in Hungarian and German. The importance of other factors, however, differs. Using parallel texts, this difference in influence can be directly compared.
- **Other strategies** 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.
- **Levels of definiteness** Clustering the levels of definiteness 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 and non-anaphoric definites, leading to a potentially better three-way distinction of levels of definiteness.
  Including more languages will yield a more fine-grained picture of cross-linguistically relevant categories of definiteness.

References I

Ariel, Mira (1988): 'Referring and Accessibility', *Journal of Linguistics* **24**(1), 65–87.

Ariel, Mira (2001): Accessibility Theory: An Overview. *In:* T. Sanders, J. Schilperoord & W. Spooren, eds, *Text Representation: Linguistic and Psycholinguistic Aspects*. Benjamins, Amsterdam, pp. 29–87.

Baayen, R. Harald, D.J Davidson & D.M Bates (2008): 'Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items', *Journal of Memory and Language* **59**(4), 390–412.

Baayen, R. Harald & Sali A. Tagliamonte (2012): 'Models, Forests and Trees of York English: Was/Were Variation as a Case Study for Statistical Practice.', *Language Variation and Change* **24**(2), 135–178.

Birner, Betty & Gregory Ward (1994): Uniqueness, Familiarity, and the Definite Article in English. In: *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society: General Session Dedicated to the Contributions of Charles J. Fillmore*. Vol. 20, BLS, pp. 93–102.

Chierchia, Gennaro (1995): *Dynamics of Meaning: Anaphora, Presupposition, and the Theory of Grammar*. University of Chicago Press, Chicago.

Clark, Herbert H. (1975): Bridging. In: *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*. TINLAP '75, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 169–174.

References II

Dryer, Matthew S. (2013): Definite Articles. *In:* M. S. Dryer & M. Haspelmath, eds, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, Matthew S (2014): 'Competing Methods for Uncovering Linguistic Diversity: The Case of Definite and Indefinite Articles (Commentary on Davis, Gillon, and Matthewson)', *Language Language* **90**(4), 232–249.

Frege, Gottlob (1892): 'Über Sinn Und Bedeutung', **100**, 25–50.

Heim, Irene (1988): *The Semantics of Definite and Indefinite Noun Phrases*. Garland Pub., New York.

Heim, Irene & Angelika Kratzer (1998): *Semantics in Generative Grammar*. Blackwell, Malden, MA.

Kamp, Hans (2002): A Theory of Truth and Semantic Representation. *In:* P. Portner & B. H. Partee, eds, *Formal Semantics*. Blackwell Publishers Ltd, pp. 189–222.

Löbner, Sebastian (1985): 'Definites', *Journal of Semantics* **4**, 279–326.

Roberts, Craige (2003): 'Uniqueness in Definite Noun Phrases', *Linguistics and Philosophy Linguistics and Philosophy* **26**(3), 287–350. OCLC: 5649366378.

Schroeder, Christoph (2011): Articles and Article Systems in Some Areas of Europe. *In:* G. Bernini & M. L. Schwartz, eds, *Pragmatic Organization of Discourse in the Languages of Europe*. Vol. 8 of *Empirical Approaches to Language Typology*, De Gruyter Mouton, Berlin, Boston, pp. 545–611.

References III

Stanley, Jason & Zoltán Gendler Szabó (2000): 'On Quantifier Domain Restriction', *Mind and Language* **15**(2), 219–261. OCLC: 359163330.

Strawson, P. F (1950): 'On Referring', *Mind* **59**(235), 320–344. OCLC: 43702178.

Thank you!