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)

Examples

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 **капетан-от**!
immediately call 3SG captain-the

ru Доложить **капитану**.
report captain.DAT

Examples

Establishing referent

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

de Darf man **den Gerüchten** trauen, Albus?
may one the rumors trust Albus

sp ¿Son ciertos **los rumores**, Albus?
are true the rumors Albus

mk Вистинити се **озборувања-та**, Албус?
true are rumors-the Albus

ru **Слухи** верны, Альбус?
rumors true Albus

Examples

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**
if you a moment time have would Mr. Cobb happily a
Jobangebot mit Ihnen besprechen.
job offer with you discuss

sp Si tienes un momento, el Sr. Cobb quiere ofrecer-te **un trabajo**.
if have.2SG a moment the Mr. Cobb wants offer-you a job

mk Ако имаш некоја минутка, г-дин Коб има **бизнис понуда** за тебе.
if have.2SG some minute Mr. Cobb has job offer for you

ru Если у тебя есть время, у мистера Кобба к тебе деловое
if PREP you is time PREP Mr. Cobb to you job
предложение.
offer

Examples

Nonspecific referent

(7) **A work placement**?

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

sp **Un internado**?
a internship

mk **Стажирање**?
internship

ru Наверное, **стажировка**?
really, internship

Data and annotation

Movies *Inception, Harry Potter 1, Pirates of the Caribbean 1, The Lion King, Lord of the Rings 1*

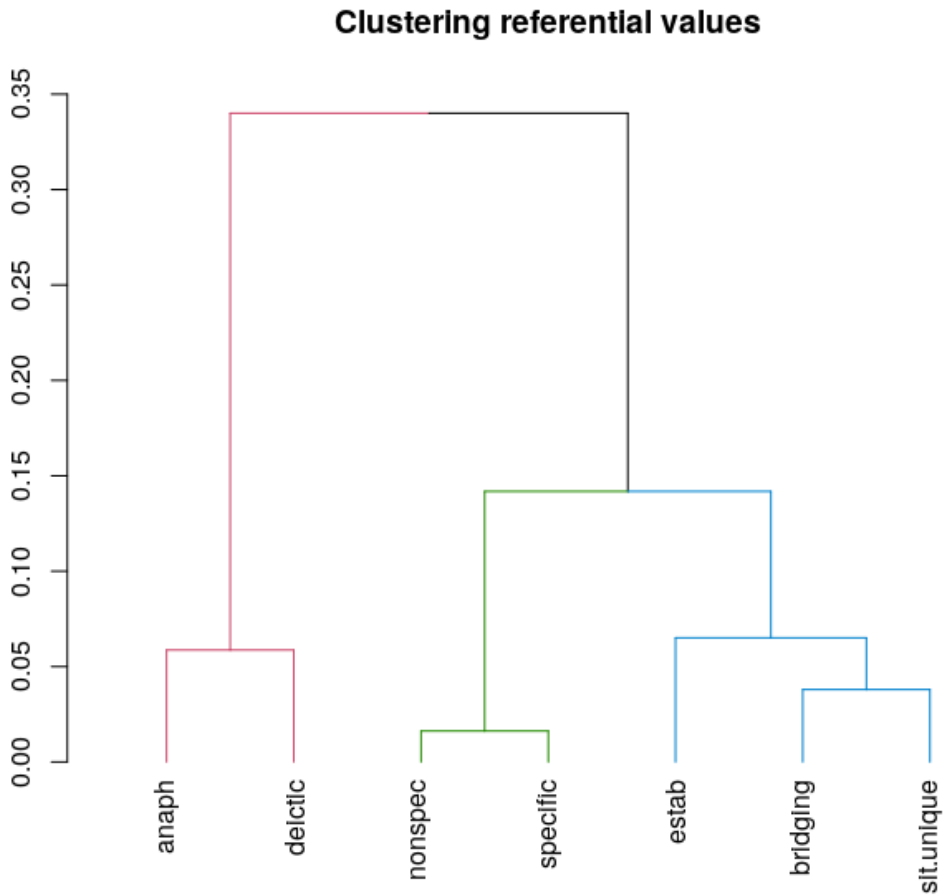
Data set 379 referential expression for each German, Spanish, Macedonian, and Russian

Annotation

- **referentiality** `deictic, anaph, sit.unique, estab, bridging, specific, nonspec`
- **syntactic function** `sbj, 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`

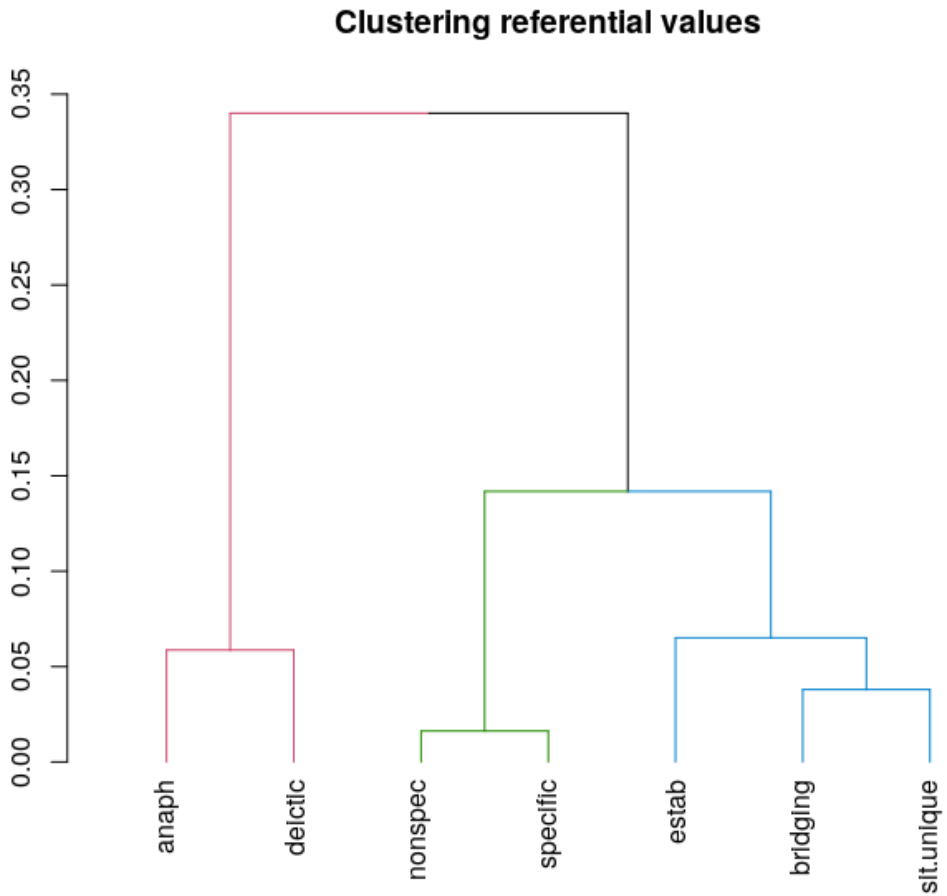
Results

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`

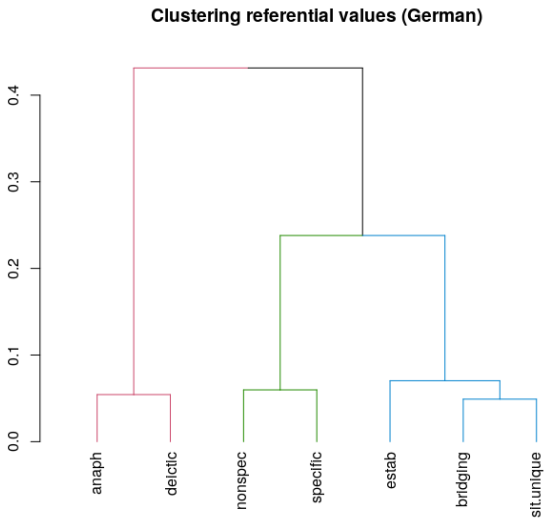
Major clusters of referential values

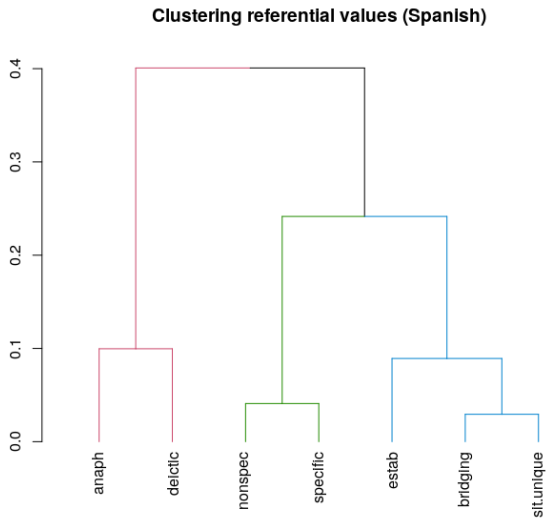


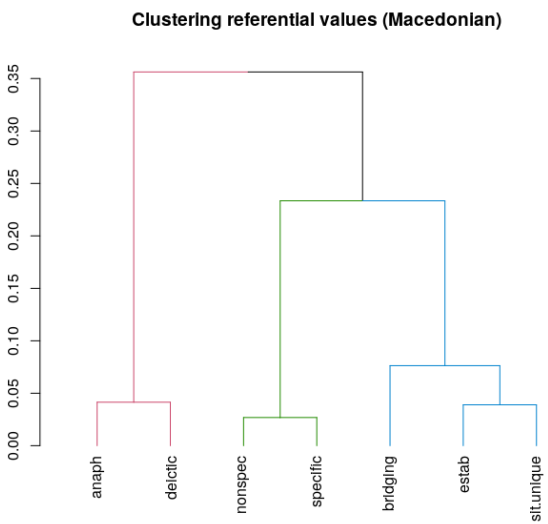
Three major referential values can be distinguished:

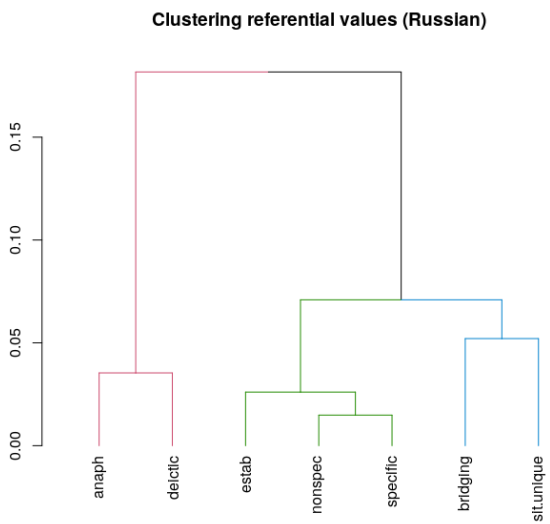
activated definite (anaphoric, deictic)
definite (establishing, bridging, situationally unique)
indefinite (specific, nonspecific)

Major clusters for each language

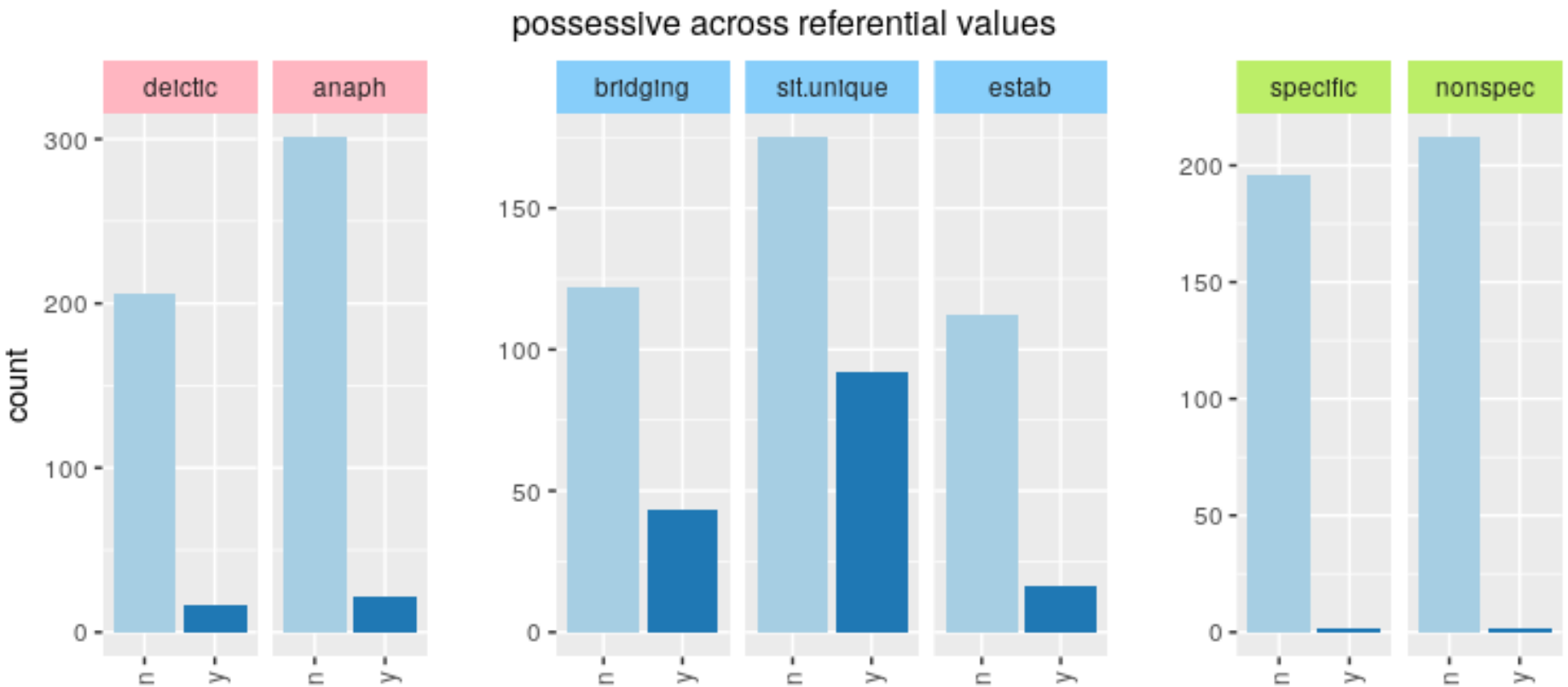




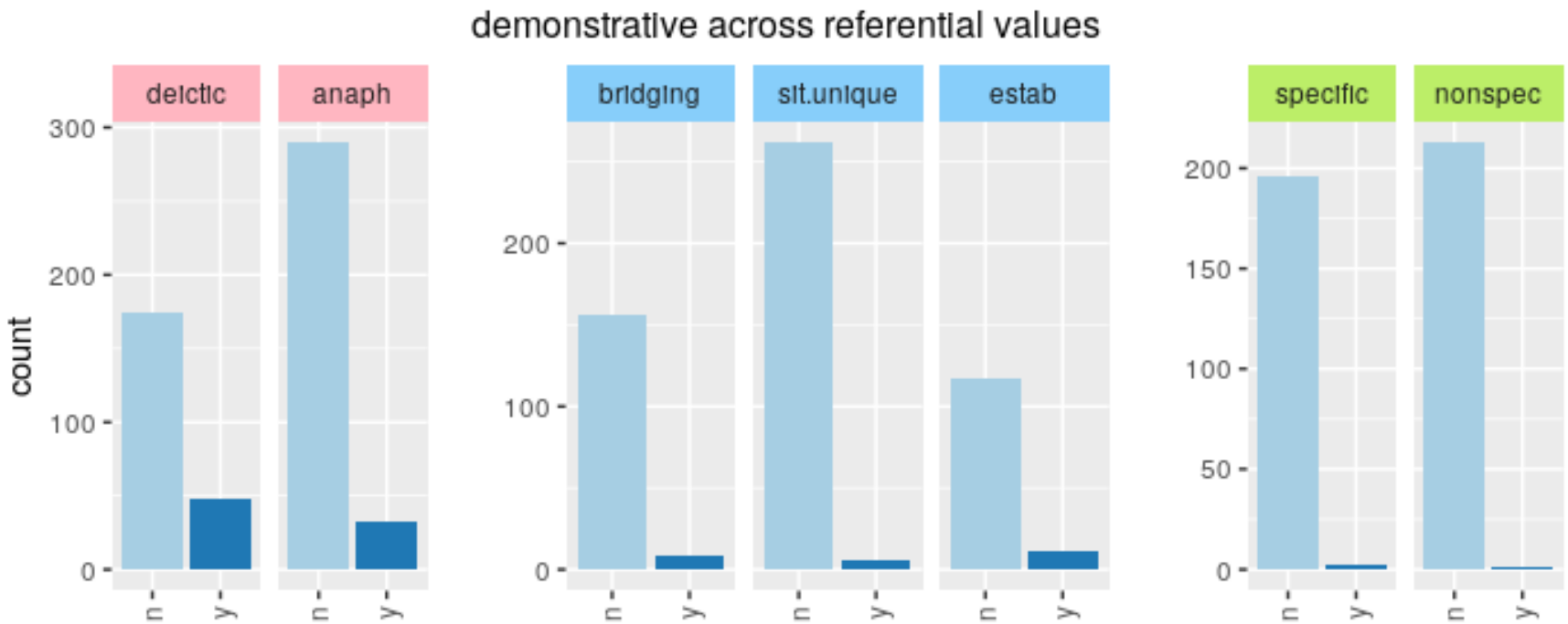




Distribution of possessives

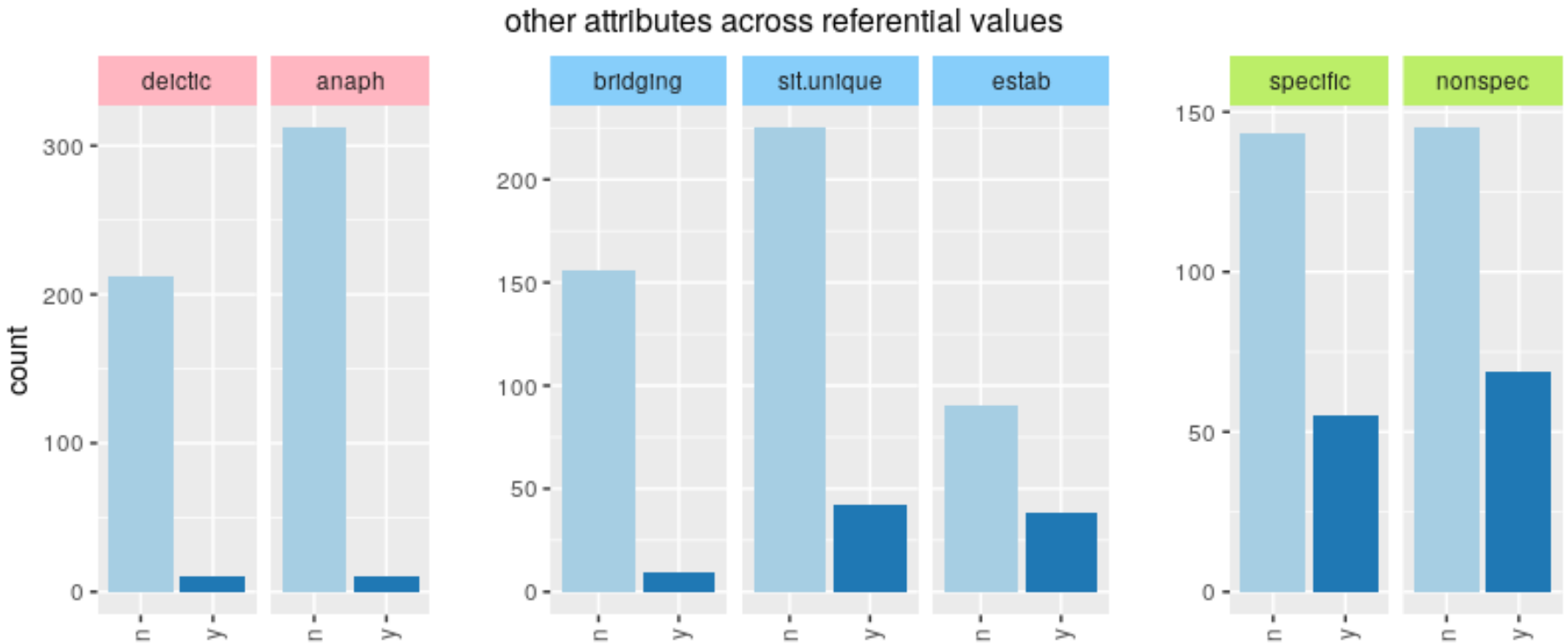


Distribution of demonstratives



Distribution of expression type

Distribution of other attributes



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)

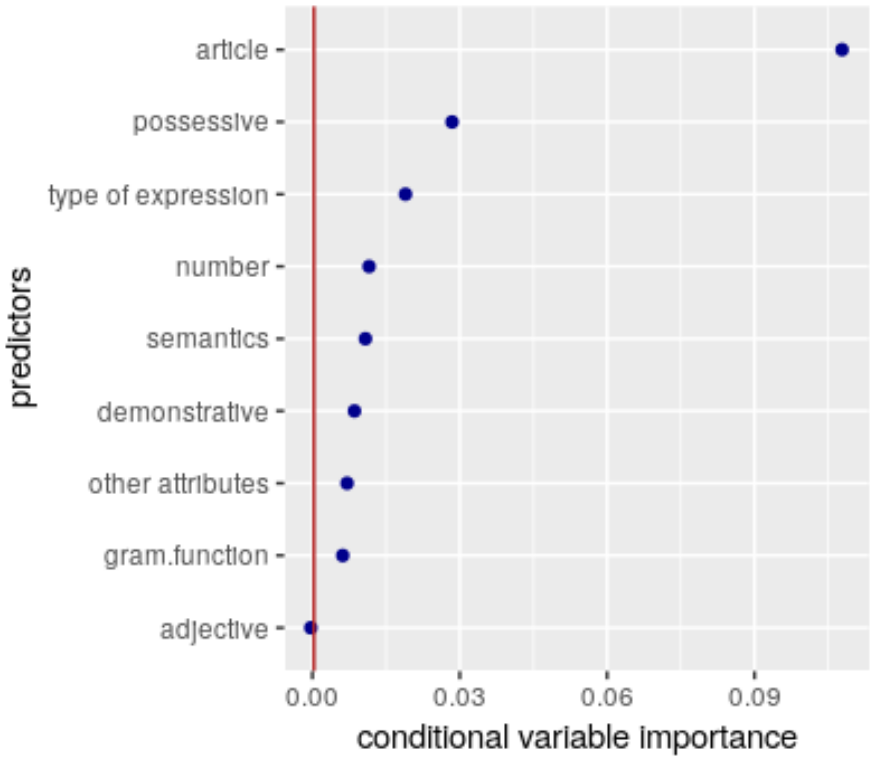
The model

```
ref ~  gram.function + semantics + expression.type +
       art + poss + dem + adj + other.attr + number
```

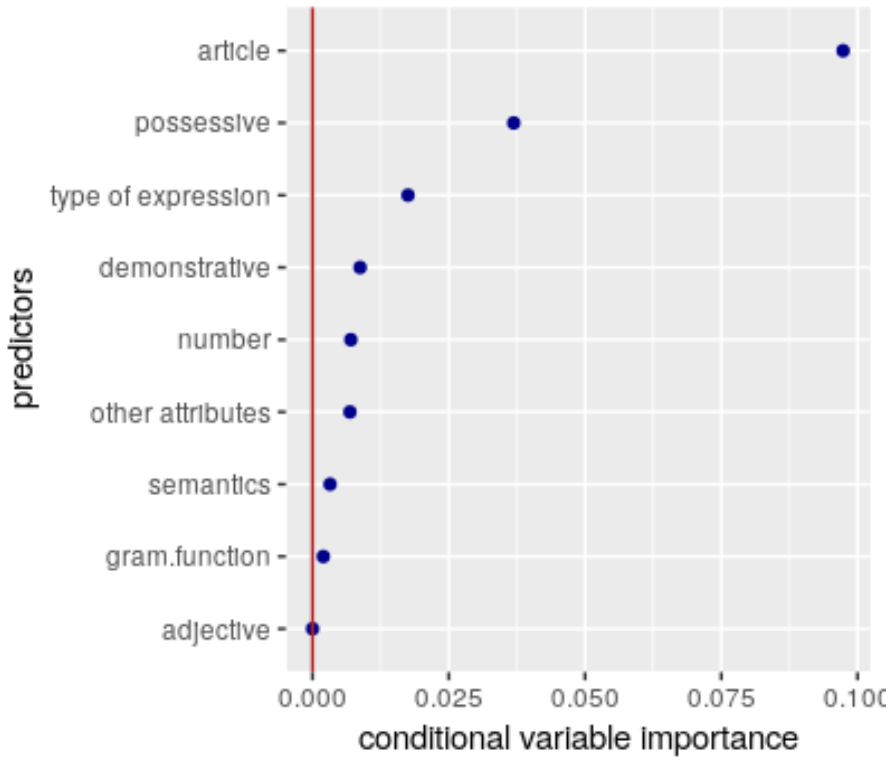
with `ref` having the three major values: `act.def`, `def`, `indef`

Variable importance

German

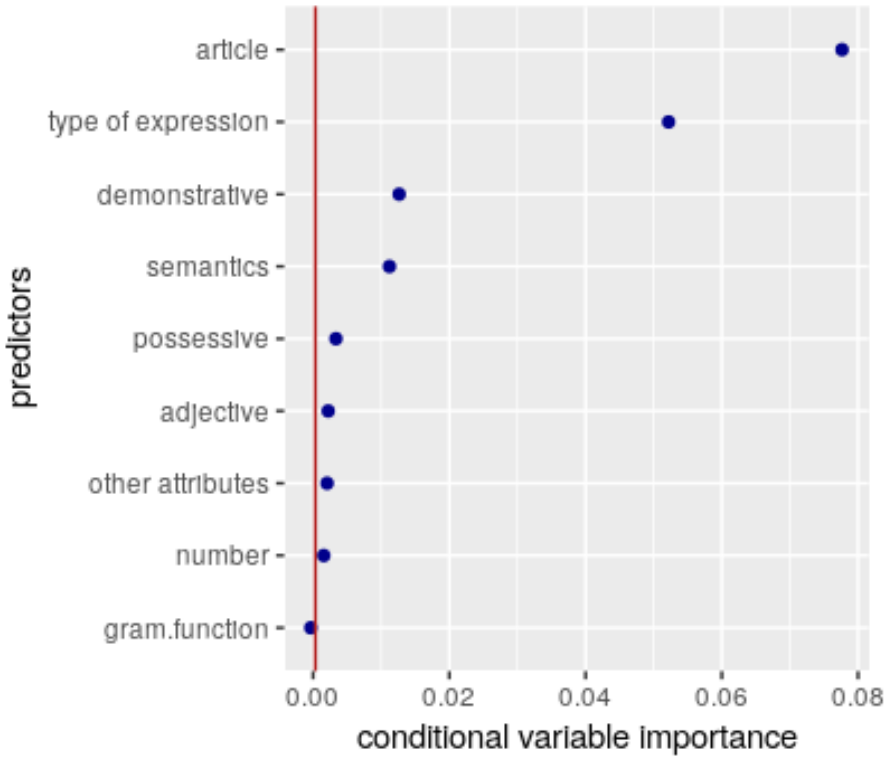


Spanish

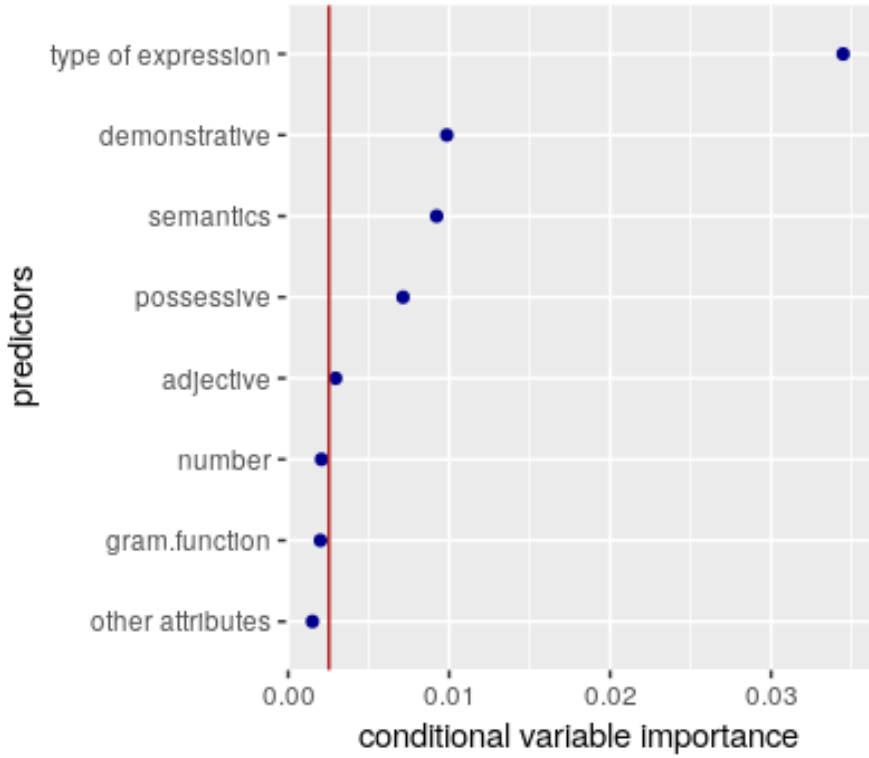


Variable importance

Macedonian



Russian



Accuracy of the Models

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

German	observed		
predicted	act-def	def	indef
act-def	**90**	14	2
def	36	**118**	3
indef	8	8	**100**

Accuracy : 0.8127
90% CI : (0.7697, 0.8507)
No Information Rate : 0.3694

Spanish	observed		
predicted	act-def	def	indef
act-def	**100**	14	5
def	29	**118**	11
indefinite	7	9	**86**

Accuracy : 0.8021
90% CI : (0.7584, 0.841)
No Information Rate : 0.372

Macedonian	observed		
predicted	act-def	def	indef
act-def	**99**	21	3
def	35	**95**	18
indef	6	23	**79**

Accuracy : 0.7203
90% CI : (0.6722, 0.7649)
No Information Rate : 0.3694

Russian	observed		
predicted	act-def	def	indef
act-def	**92**	18	8
def	30	**87**	21
indefinite	12	35	**76**

Accuracy : 0.6728
90% CI : (0.6231, 0.7199)
No Information Rate : 0.3694

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.

German	observed		
predicted	act-def	def	indef
act-def	**90**	14	4
def	28	**92**	21
indef	16	34	**80**

Accuracy :0.6913
90% CI : (0.6421, 0.7375)
No Information Rate : 0.3694

Spanish	observed		
predicted	act-def	def	indef
act-def	**99**	15	11
def	23	**98**	17
indefinite	14	28	**74**

Accuracy : 0.715
90% CI : (0.6667, 0.76)
No Information Rate : 0.372

Macedonian	observed		
predicted	act-def	def	indef
act-def	**95**	18	8
def	33	**75**	21
indef	12	43	**71**

Accuracy : 0.6359
90% CI : (0.5852, 0.6844)
No Information Rate : 0.3694

Russian	observed		
predicted	act-def	def	indef
act-def	**92**	18	8
def	30	87	21
indefinite	12	35	**76**

Accuracy : 0.6728
90% CI : (0.6231, 0.7199)
No Information Rate : 0.3694

Summing up

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!

References

Ariel, Mira (1988). "Referring and Accessibility". In: *Journal of Linguistics* 24.1, pp. 65–87.

Becker, Laura (2018). "Articles in the World's Languages". PhD thesis. University of Leipzig.

Dryer, Matthew S. (2014). "Competing Methods for Uncovering Linguistic Diversity: The Case of Definite and Indefinite Articles (Commentary on Davis, Gillon, and Matthewson)". In: *Language Language* 90.4, pp. 232–249.

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski (1993). "Cognitive Status and the Form of Referring Expressions in Discourse". In: *Language* 69.2, pp. 274–307.

Hawkins, John A. (1978). *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. London: Croom Helm Humanities Press.

Himmelmann, Nikolaus P. (1997). *Deiktikon, Artikel, Nominalphrase: Zur Emergenz Syntaktischer Struktur*. Tübingen: Niemeyer.

Strobl, Carolin, James Malley, and Gerhard Tutz (2009). "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests". In: *Psychological methods* 14, pp. 323–48.

Examples

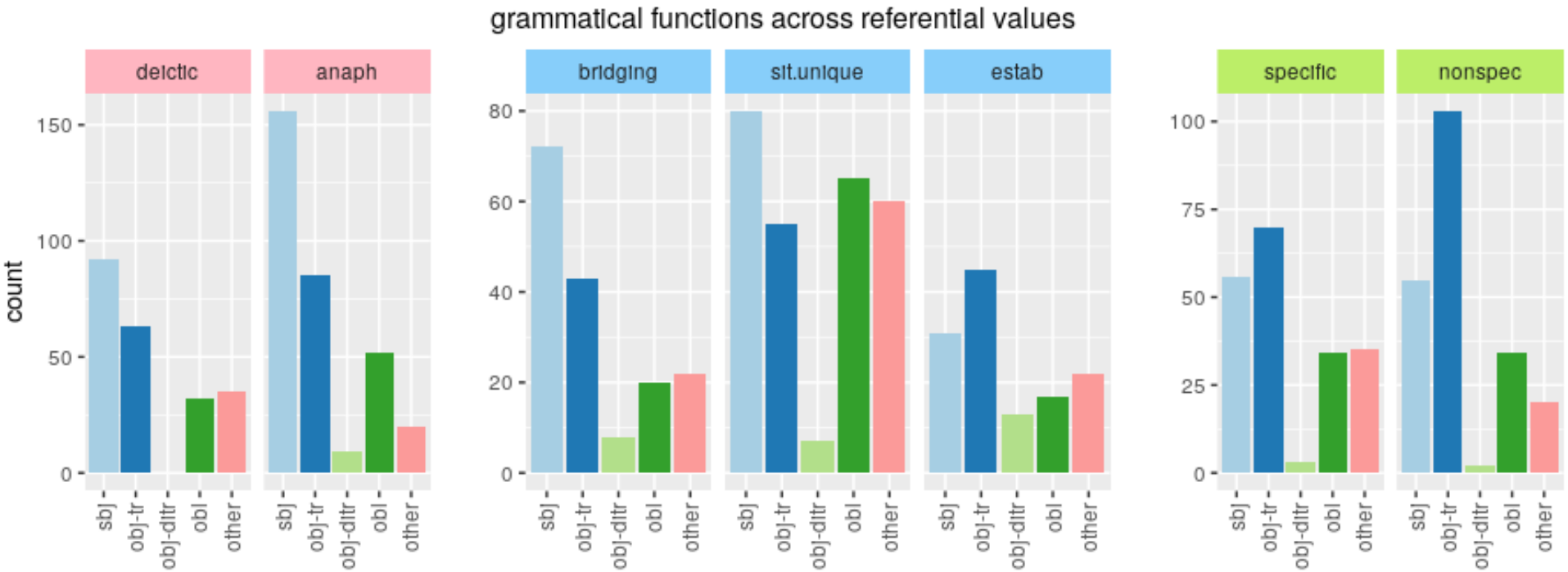
Even though the referential expressions are equivalent, they do not have to have the identical referential value in all languages:

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

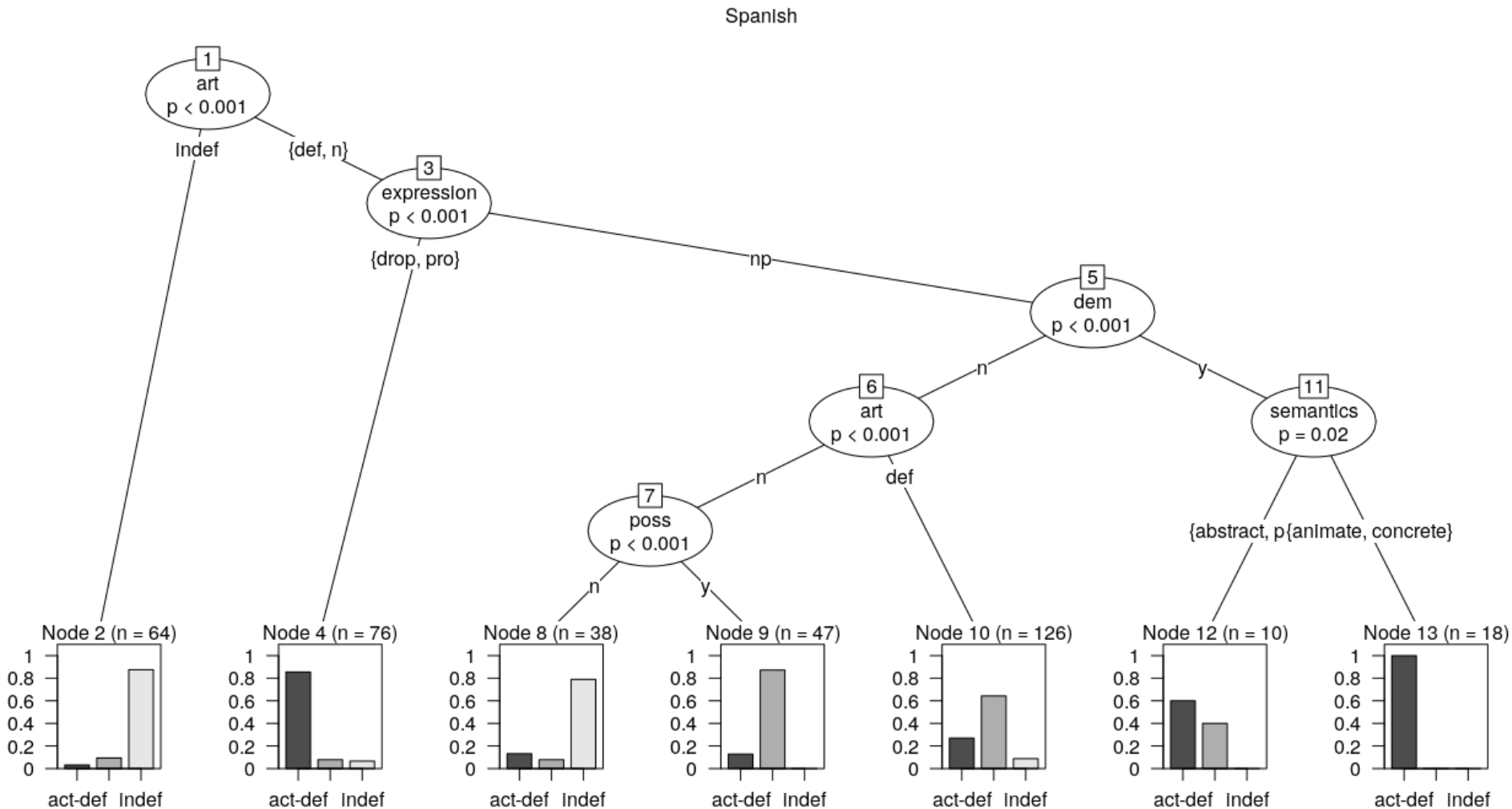
sp a. Y en **la oscuridad** de la cueva de Gollum, aguardó.
and in the darkness of the cave of Gollum, waited.3SG
b. **La oscuridad** se empezó a filtrar en los bosques del
the darkness REFL started.3SG PREP infiltrate in the forests of.the
mundo.
world

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

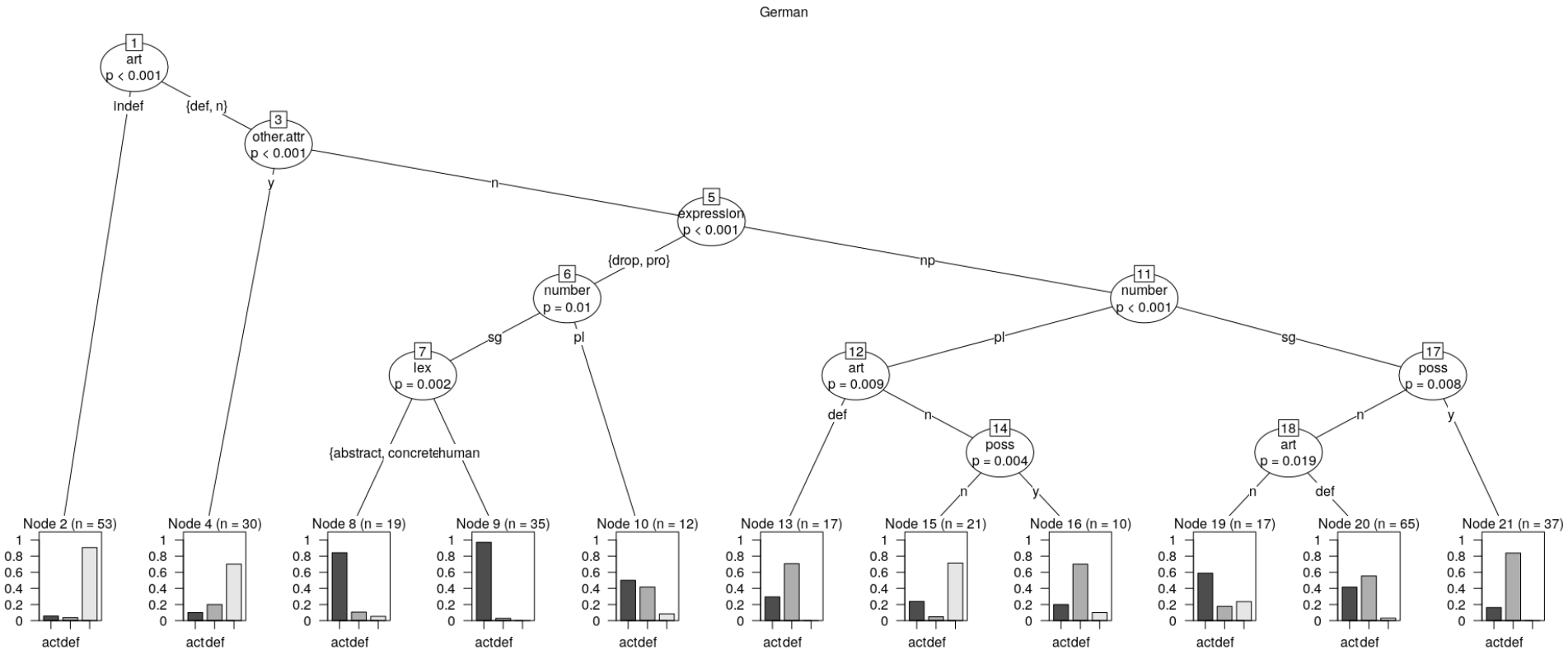
Distribution of grammatical functions



Decision tree (Spanish)



Classification tree (German)



Classification tree (Russian)