SLE 2018 | WS Comparative corpus linguistics

Is intransitive subject the preferred role for introducing new referents?

Evidence from corpus-based typology

Stefan Schnell
ARC CoEDL / U Melbourne

Nils Norman Schiborr
University of Bamberg

Geoffrey Haig
University of Bamberg

01 September 2018
v1.0

DFG

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

THE UNIVERSITY OF
MELBOURNE

UNIVERSITY OF BAMBERG
OTTO-FRIEDRICH-UNIVERSITÄT BAMBERG

Overview

1. Introduction:
   Argument structure as architecture for referent introduction
2. Language corpora, annotations, and methods
3. Findings:
   P arguments as the predominant entry point for new referents
4. Conclusions and outlook

Once upon a time...

(1) Grimm's *Fairy tales*, "The wolf and the seven young kids"

a. *There was once upon a time [an old goat]$_{S}$*

b. *[who]$_{A}$ had [seven little kids]$_{P}$*

c. *and Ø$_{A}$ loved [them]$_{P}$ . . .*

(2) *Once upon a time [an old goat]$_{A}$ had [seven kids]$_{P}$ . . .*

S and P, vs. A

… most lexical mentions occur in absolutive argument positions (S or O [= P]), but are avoided in the ergative (A) slot, which is mostly restricted to reduced forms (pronoun, agreement, zero). Correspondingly, most new mentions occur in S or O [= P], with few occurring in A.

– Du Bois 2017: 29, emph. added

Preferred argument structure

- Du Bois' (1987, 2003, 2017) '**preferred argument structure**':
  - certain syntactic roles are systematically associated with particular information status
  - due to assumed constraints on information processing

    Chafe's (1987) 'one new concept at a time constraint'

    Lambrecht's (1994) 'separation of role and reference'

- cf. 'pragmatic linking' (Durie 2003):

  S + P provide "a predictable locus for unpredictable work"

  (Du Bois 2003a: 47)

S and P as vectors of new information

Among the things that a speaker may know about the verb come in, *for example, is that its S role provides a reliably usable slot for introducing a new human protagonist into a discourse. Likewise, the O [= P] of the verb* meet *may serve a similar function.*

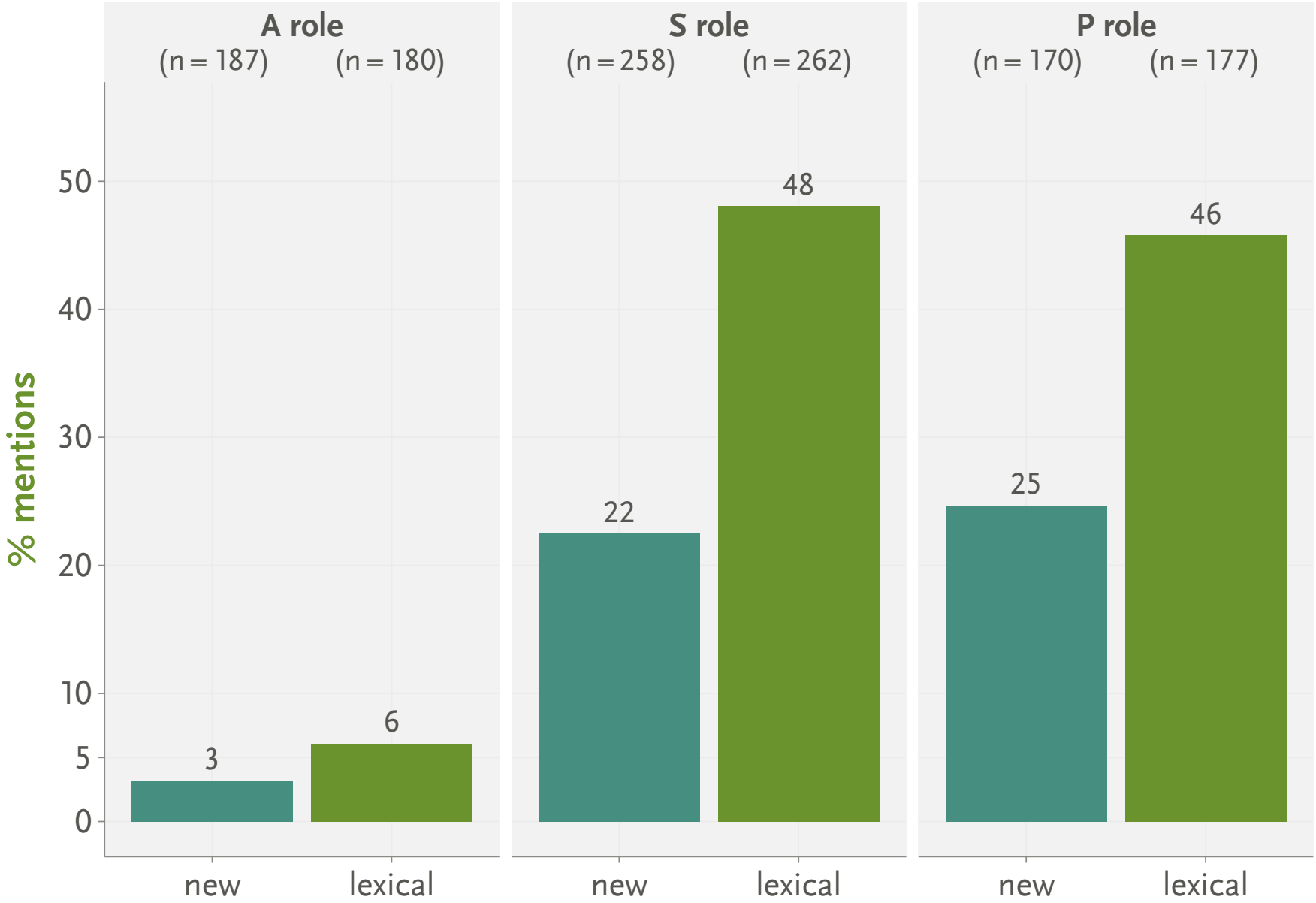
– Du Bois 2003a: 40, emph. added

Cognitive constraints

The way to fulfilment [i.e. to abiding by the constraints on argument structure in discourse processing] is via a simple principle of discourse: Speakers need not say everything in one clause. Facing cognitive constraints that could frustrate their expressive goals, speakers can simply mobilize their planning capacity to organize a series of successive clauses.

– Du Bois 2003b: 76, emph. added

Sakapultek (Mayan, Du Bois 1987: 822; 828)



Intransitive introduction

- **but,**
  it is **intransitive clauses** that are added
  for the sake of information flow

  rather than for their "conceptual content
  or semantic one-placeness" (Du Bois 1987: 831)
- e.g. predicates like *arrive, appear, come in*

- hence '**intransitive introduction, transitive narration**':
  speakers prefer the S role for introducing new referents
  to minimize processing costs

Intransitive introduction

(3) *Pear stories* (Chafe 1980)

a. *a man was picking pears* (p. 316, speaker 16)

b. *there was a man, who was picking pears* (p. 306, speaker 7)

- both options are grammatical,
- but the second is **preferred**

Research questions

1. Is the **S role**, as claimed, cross-linguistically "specialized" for the **introduction of new referents**, regardless of language type?

Research questions

1. Is the **S role**, as claimed, cross-linguistically "specialized" for the **introduction of new referents**, regardless of language type?
2. In what ways can this be meaningfully evaluated from a **statistical perspective**?

Research questions

1. Is the **S role**, as claimed, cross-linguistically "specialized" for the **introduction of new referents**, regardless of language type?
2. In what ways can this be meaningfully evaluated from a **statistical perspective**?
3. Are the **S and P roles** comparable in this respect?

Research questions

1. Is the **S role**, as claimed, cross-linguistically "specialized" for the **introduction of new referents**, regardless of language type?
2. In what ways can this be meaningfully evaluated from a **statistical perspective**?
3. Are the **S and P roles** comparable in this respect?
4. What is the role of non-core arguments in managing new information?

The sample

- non-elicited, monologic spoken narratives
  from the freely accessible **Multi-CAST collection**
  (Haig & Schnell 2015)

corpus	affil.	n(clauses)	citation
Cypriot Greek	I.-E.	1 071	Vollmer & Hadjidas 2015
English	I.-E.	1 244	Schiborr 2015
North. Kurdish	I.-E.	1 389	Haig & Thiele 2015 *
Sanzhi Dargwa	Nakh-Dagh.	1 702	Forker & Schiborr in prog.
Teop	Oceanic	1 272	Mosel & Schnell 2015
Vera'a	Oceanic	2 377	Schnell 2015
totals		**8 971**	

* with additional contributions by Maria Vollmer

Methodology: annotations

- the corpora have been manually annotated for
  - the form and **role** of referring expressions,
    (with GRAID, Haig & Schnell 2014)
  - the **identity** of each mention of a referent, and
    (with RefIND, Schiborr et al. 2018)
  - the **information status** of new referents
    (with RefLex, Riester & Baumann 2017)

Examples

(4) **Sanzhi Dargwa** [sanzhi_devil_034]

xun-ne-b		*suk*	*b-ič-ib*	*k:urt:a*
road-SPR-N	Ø	meet	N-occur.PFV-PRET	fox
np:l	0.h:s	other	v:pred	np.d:p
	0002			0031
				new

'On the road (he) met a fox.'

GRAID: grammatical relations

- defined as per Andrews (2007, 1985) as
  '**generalizations of semantics prototype roles across encoding properties**'
  - **S** = subjects of intransitive clauses
  - **A** = function coded like prototypical agents
  - **P** = function coded like prototypical patients

GRAID: grammatical relations

- additionally, we distinguish
  - goals, recipients, and addressees,
  - static locations,
  - other oblique arguments, and
  - others (possessors, predicates, direct address, etc.)

RefIND: identifying discourse referents

- **discourse referents**
  - are linguistic representations of construed entities,
  - time-stable within the universe of discourse and
  - trackable across states-of-affairs throughout a discourse (Du Bois 1980); but which
  - exclude various instances of nominal expressions,
    - e.g. those under scope of negation,
      predicates of classificatory clauses,
      conflated objects, etc.

RefIND: new vs. given

- at the first mention in linear order of a particular referent, that referent is considered **newly introduced**;
- all subsequent mentions of the same referent are assumed to be **given**

RefLex: new, unused, and bridging

bridging

- **a.** referent inferable from frame semantics or
- **b.** a previously mentioned situation, or
- **c.** anchored to an already given referent

unused

a globally known entity,
via encyclopædic or cultural knowledge

(brand) new

referent not otherwise inferable or known

Examples

(5) **Sanzhi Dargwa** [sanzhi_devil_038]

kːurtːa-l	*b-ičː-ib*	*hel-i-j*	*cin-na*	*ʁez*
fox-ERG	N-give.PFV-PRET	that-OBL-DAT	REFL.SG-GEN	hair
np.d:a	v:pred	pro.h:g	ln_refl.d:poss	np:p
0031		0002	0031	0032
				bridging

‘The fox gave him one of its hairs.’

Methodology: automation

- based on the annotations, we can algorithmically determine
  - the **frequency** of each unique referent,
  - the **position** of each mention relative to others
    (e.g. for newness, lookback distance, etc.), and
  - the **relative proportions** of each group
    (i.e. `corpus | role | information status`)

Methodology: procedure

1. the texts are annotated in ELAN,* stored as XML files
2. a custom R script reads these XML files,
   and transforms them into a table
   (data and scripts are available in the `multicastR` package**)
3. each row of the table represents one 'grammatical word',
   i.e. the smallest GRAID annotation unit
4. the table is filtered for the heads of referring expressions,
   thereby excluding all non-referring material,

   so that one data point = one referring expression

* https://tla.mpi.nl/tools/tla-tools/elan/
** https://CRAN.R-project.org/package=multicastR

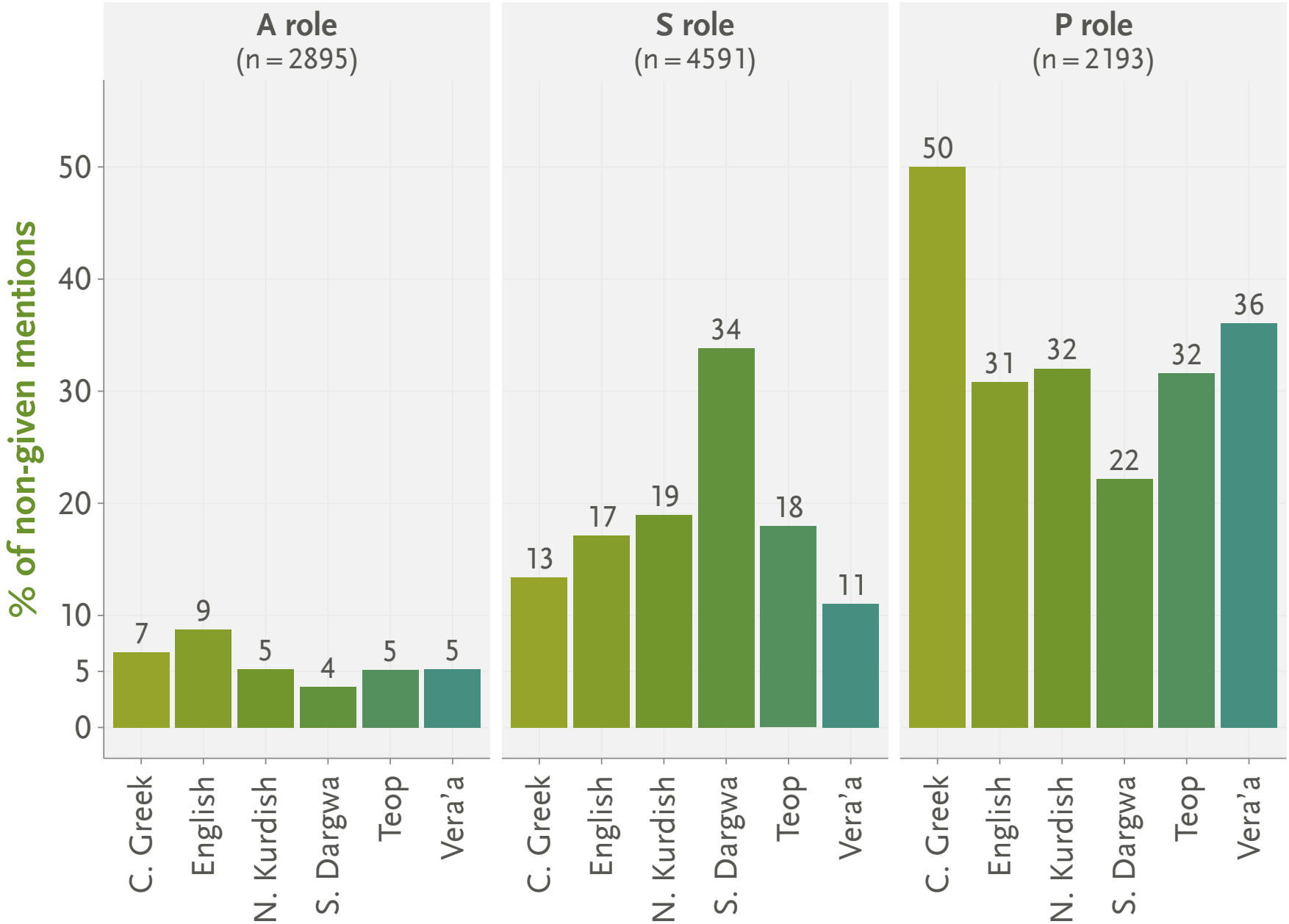
Given and new: proportions

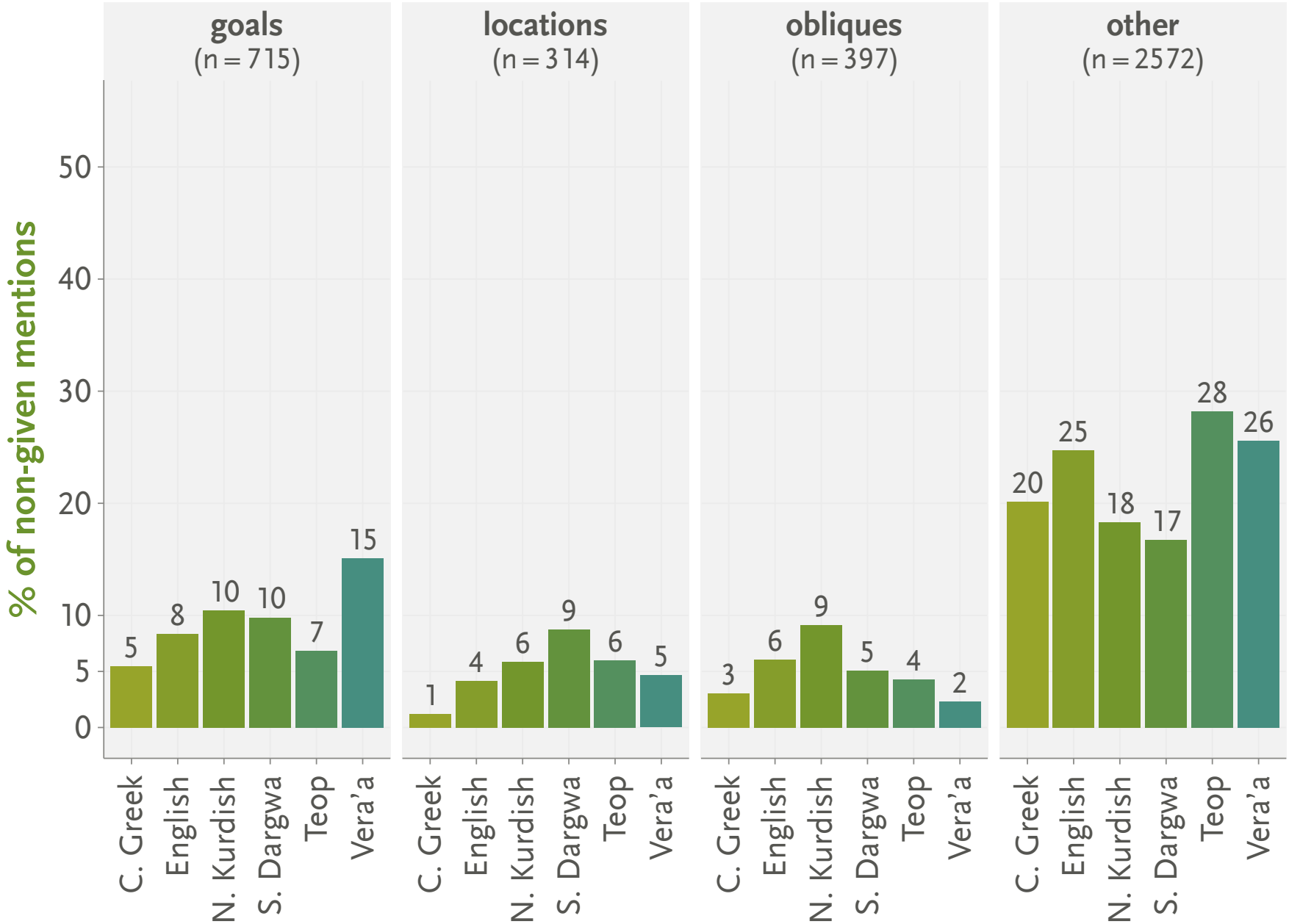
- tracked referents 1 273 across 29 texts
- **total mentions** **13 677**
  - brand new 531 (4 %)
  - bridging 565 (4 %)
  - given 12 533 (92 %)

- ‘tracked’ referent
  = unique referent with at least two mentions
  (excluding 1104 referents mentioned only once)

New mentions (A)

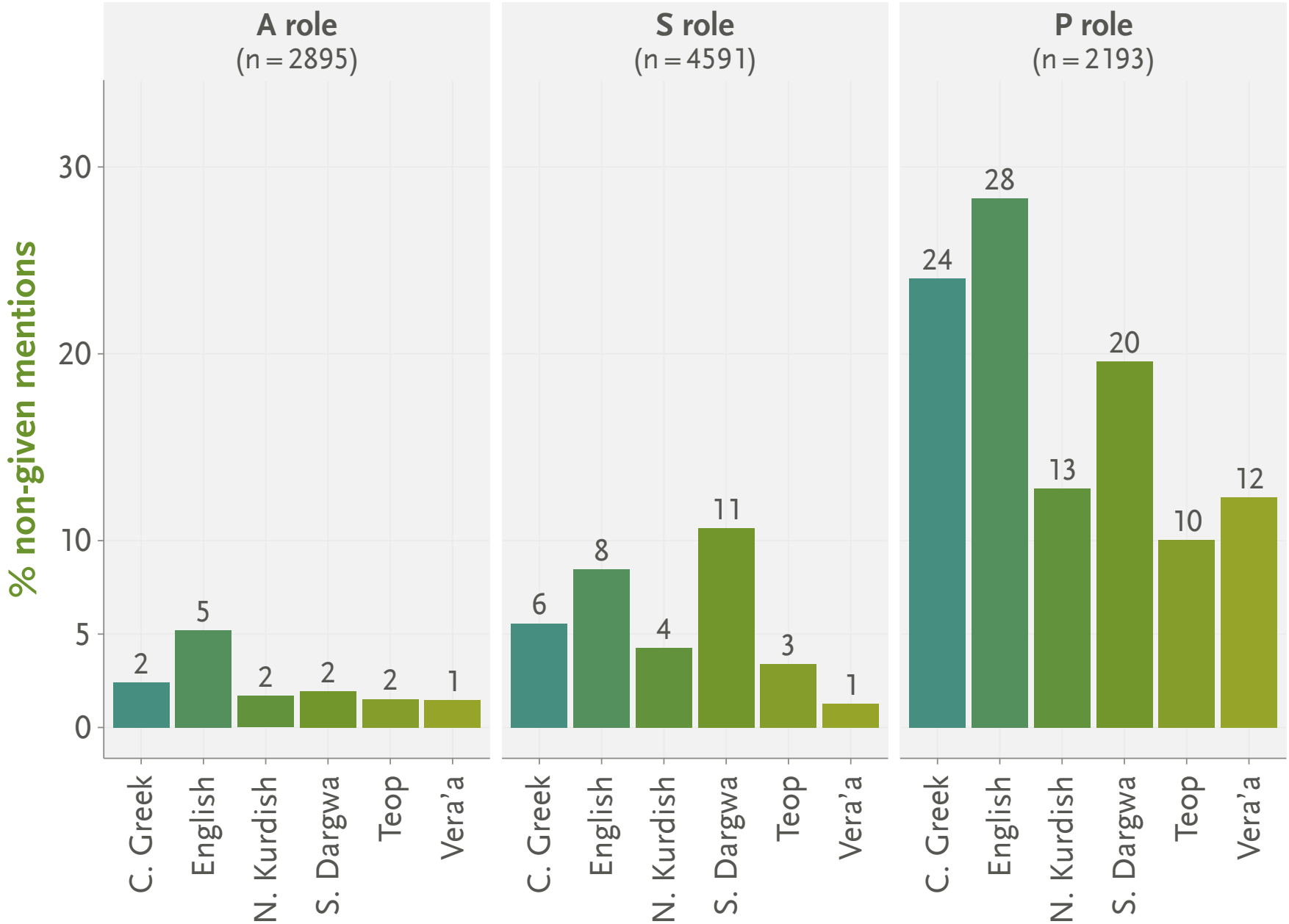
given a newly introduced referent,
what is the probability of it being in a particular role?

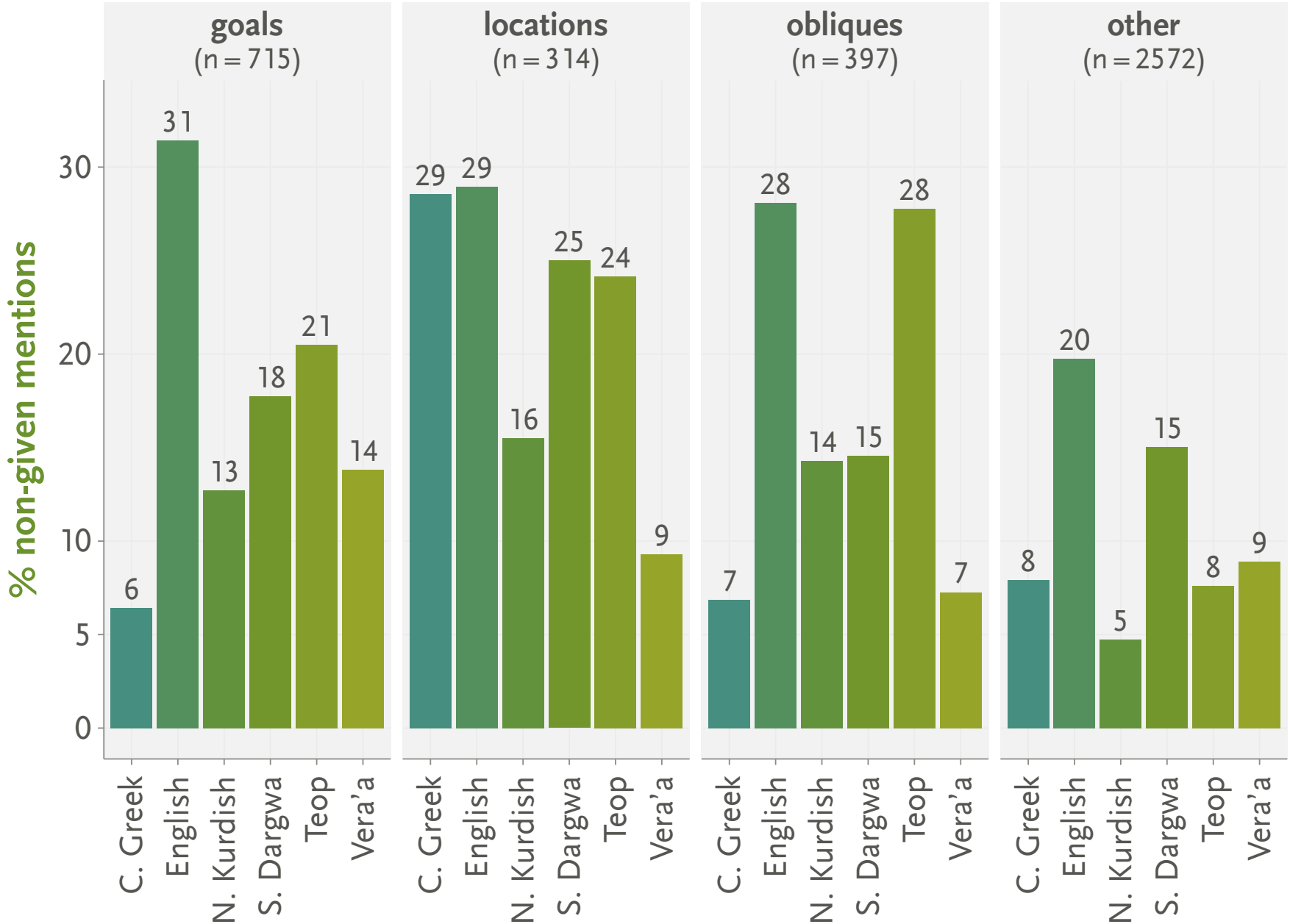

A role
(n = 2895)
S role
(n = 4591)
P role
(n = 2193)
% of non-given mentions
0
5
10
20
30
40
50
7
9
5
4
5
5
13
17
19
34
18
11
50
31
32
22
32
36
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


goals
(n = 715)
locations
(n = 314)
obliques
(n = 397)
other
(n = 2572)
% of non-given mentions
50
40
30
20
10
5
0
5
8
10
10
7
15
1
4
6
9
6
5
3
6
9
5
4
2
20
25
18
17
28
26
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a

New mentions (B)

what proportion of each role is dedicated to newly introduced referents?


A role
(n = 2895)
S role
(n = 4591)
P role
(n = 2193)
% non-given mentions
0
5
10
20
30
2
5
2
2
2
1
6
8
4
11
3
1
24
28
13
20
10
12
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


goals
(n = 715)
locations
(n = 314)
obliques
(n = 397)
other
(n = 2572)
% non-given mentions
0
5
10
20
30
6
31
13
18
21
14
29
29
16
25
24
9
7
28
14
15
28
7
8
20
5
15
8
9
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a

The prominence of P

- consistently across all corpora,

  $p(\mathbf{P}_{\mathbf{new}}) > p(\mathbf{S}_{\mathbf{new}}) > p(\mathbf{A}_{\mathbf{new}})$

  (Fisher's exact tests for S and P yield $p < 0.0001$ for all corpora)

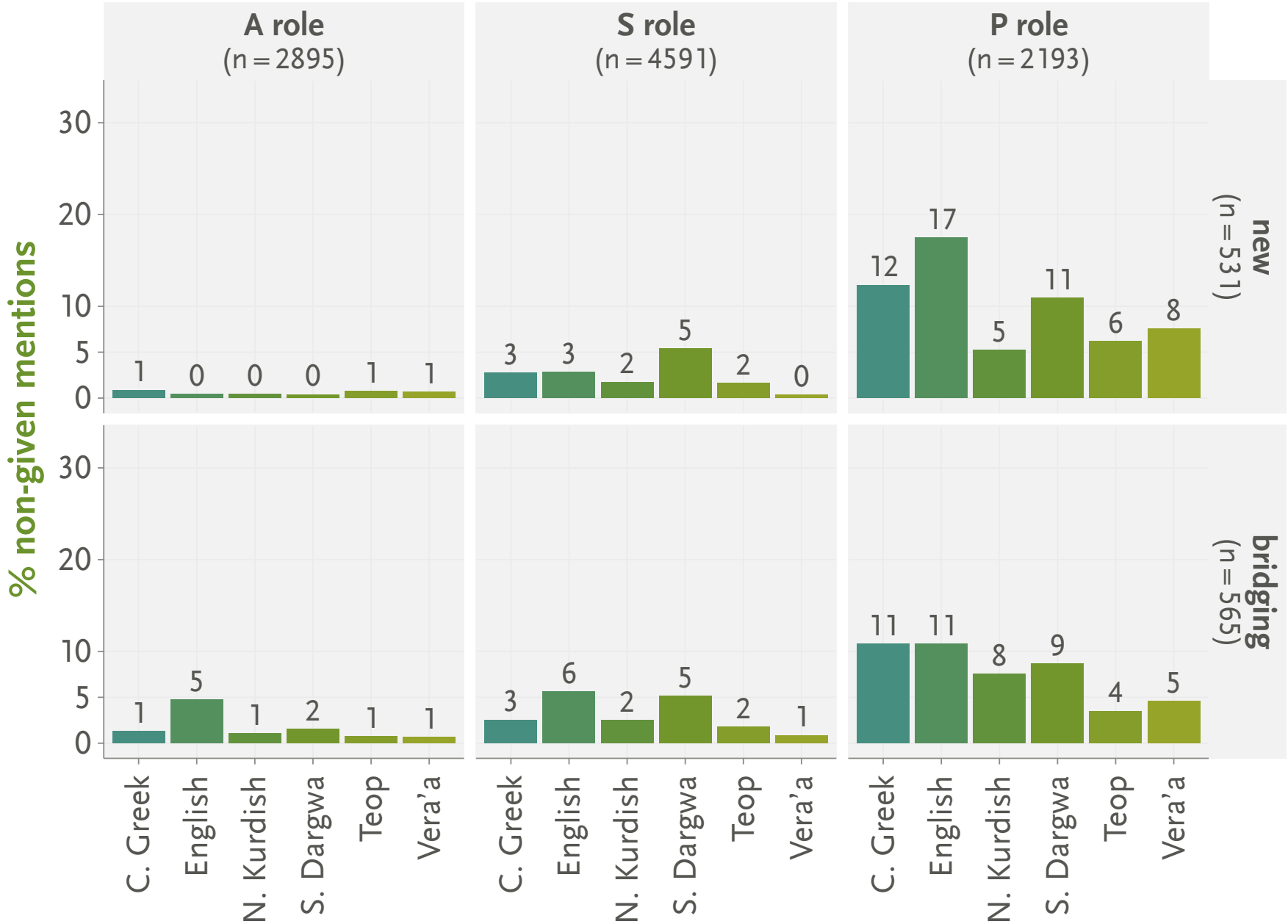
- in general, **the P role**
  tends to have the largest fraction of mentions of
  any core argument dedicated to new introductions

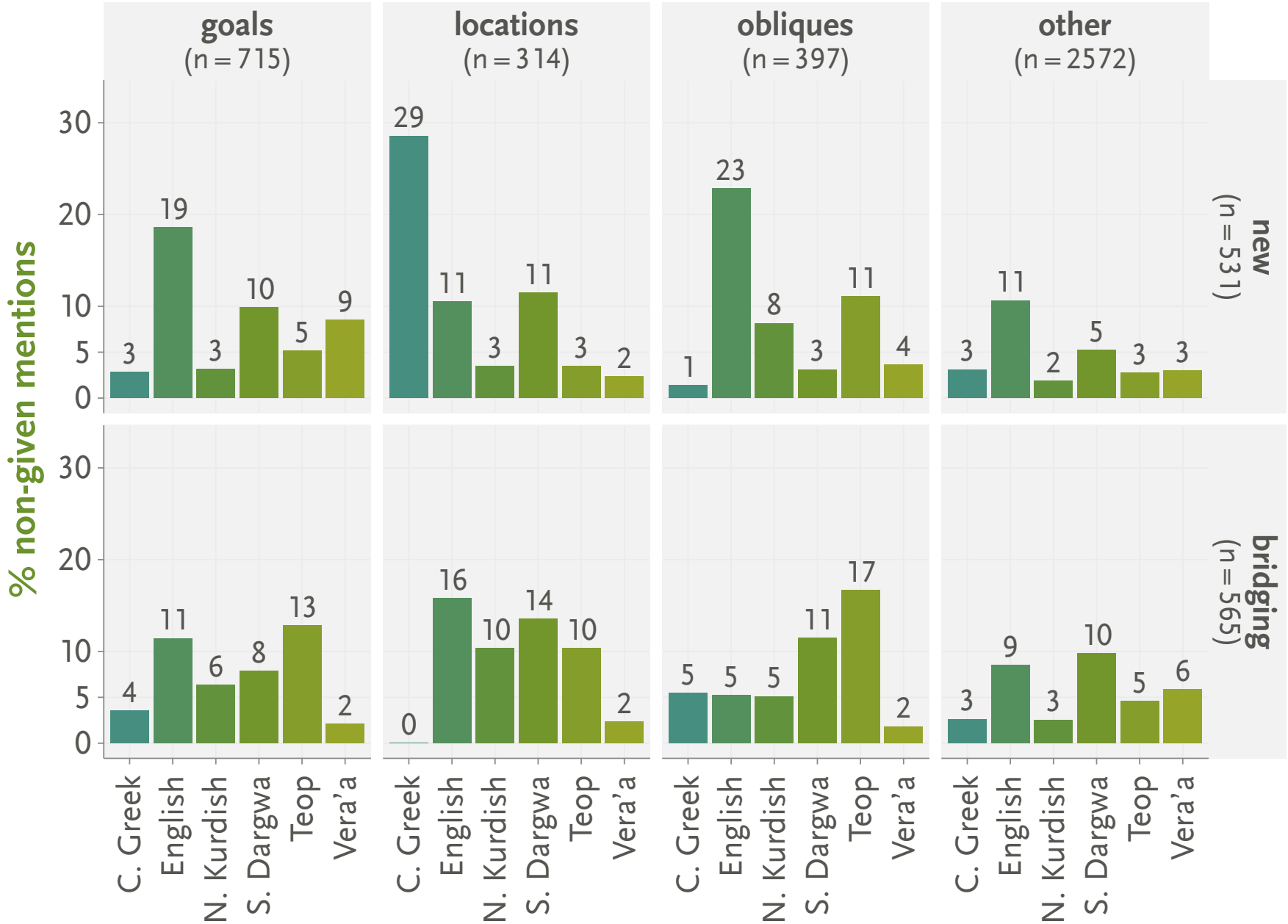
New vs. bridging

- discourse-new referents that are **evoked/inferable (bridging)**
  should be easier to process, hence
  require less effort to place into predictable loci
- generally, **bridging** should thus
  - be more freely associated with the constraint role of A,
  - and likely less frequently in P and
    the processing-aiding position S,
  - than (brand) new referents

New vs. bridging

what proportion of each role is dedicated to brand new and bridging introductions?


A role
(n = 2895)
S role
(n = 4591)
P role
(n = 2193)
new
(n = 531)
bridging
(n = 565)
% non-given mentions
0
5
10
20
30
1 0 0 0 1 1
3 3 2 5 2 0
12 17 5 11 6 8
1 5 1 2 1 1
3 6 2 5 2 1
11 11 8 9 4 5
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a


goals
(n = 715)
locations
(n = 314)
obliques
(n = 397)
other
(n = 2572)
new
(n = 531)
bridging
(n = 565)
% non-given mentions
0
5
10
20
30
3
19
3
10
5
9
29
11
3
11
3
2
1
23
8
3
11
4
3
11
2
5
3
3
4
11
8
13
2
0
16
10
14
10
2
5
5
5
11
17
2
3
9
3
10
5
6
C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera'a

Conclusion: S and new information

- the S role appears "specialized" for new introductions
  only by the **proportion** of S among all new introductions

Conclusion: S and new information

- the S role appears "specialized" for new introductions
  only by the **proportion** of S among all new introductions

- **but,**
  S arguments are overall highly frequent,
  hence offering more potential 'landing sites'
  for new information

Conclusion: S and new information

- the S role appears "specialized" for new introductions
  only by the **proportion** of S among all new introductions

- **but,**
  S arguments are overall highly frequent,
  hence offering more potential 'landing sites'
  for new information

- from **an intra-role perspective**,
  new mentions make up only a small fraction
  of the S arguments in a text

Conclusion: P and new information

- **instead,**
  the **P role** (and certain non-core arguments)
  harbour much **larger proportions of new mentions**

- for the **P role** especially,
  we might hypothesize a genuine cross-linguistic association with new information

Conclusion: P and new information

- this association may in turn be motivated
  by **the association of new mentions with certain semantic roles**
  (rather than by pragmatic linking to a syntactic position)

- **also:**
  linking of a new referent **to an already established one** in a transitive construction

Conclusion: non-core functions

- **non-core functions** generally harbour
  high proportions of new mentions
- goals/recipients/addressees tend to bear fewer
  new mentions (especially when human)
- than locations and other oblique arguments
  (in particular when non-human)

- **however,**
  large inter-corpus variability within these roles
  (possibly more content-sensitive?)

Conclusions: new vs. bridging

- **no clear effect**
- **but again,**
  it is **P**, but not S, that is most
  inclined to host brand-new mentions!

Corpus-based typology

- **challenges**:
  - size of data sets and representativeness
  - comparability of data sets
  - inter-annotator differences
  - open availability of data and methods
  - also: annotation workload

Going forward

- **with our established methods,**
  we can readily determine anaphoric relations
  (e.g. lookback distances, role continuity)
  as well as related anaphoric forms
  and their syntactic functions,
  and more

Going forward

- furthermore, a **notion of 'topic'** can be secondarily derived,
- e.g. as **expressions** that
  - **a.** occur above a certain frequency threshold;
  - **b.** occur in S or A role in strings of consecutive clauses;
  - **c.** are most frequently realized via reduced forms (pronouns, zero);

    etc.

all data will (in the near future) be freely accessible online at

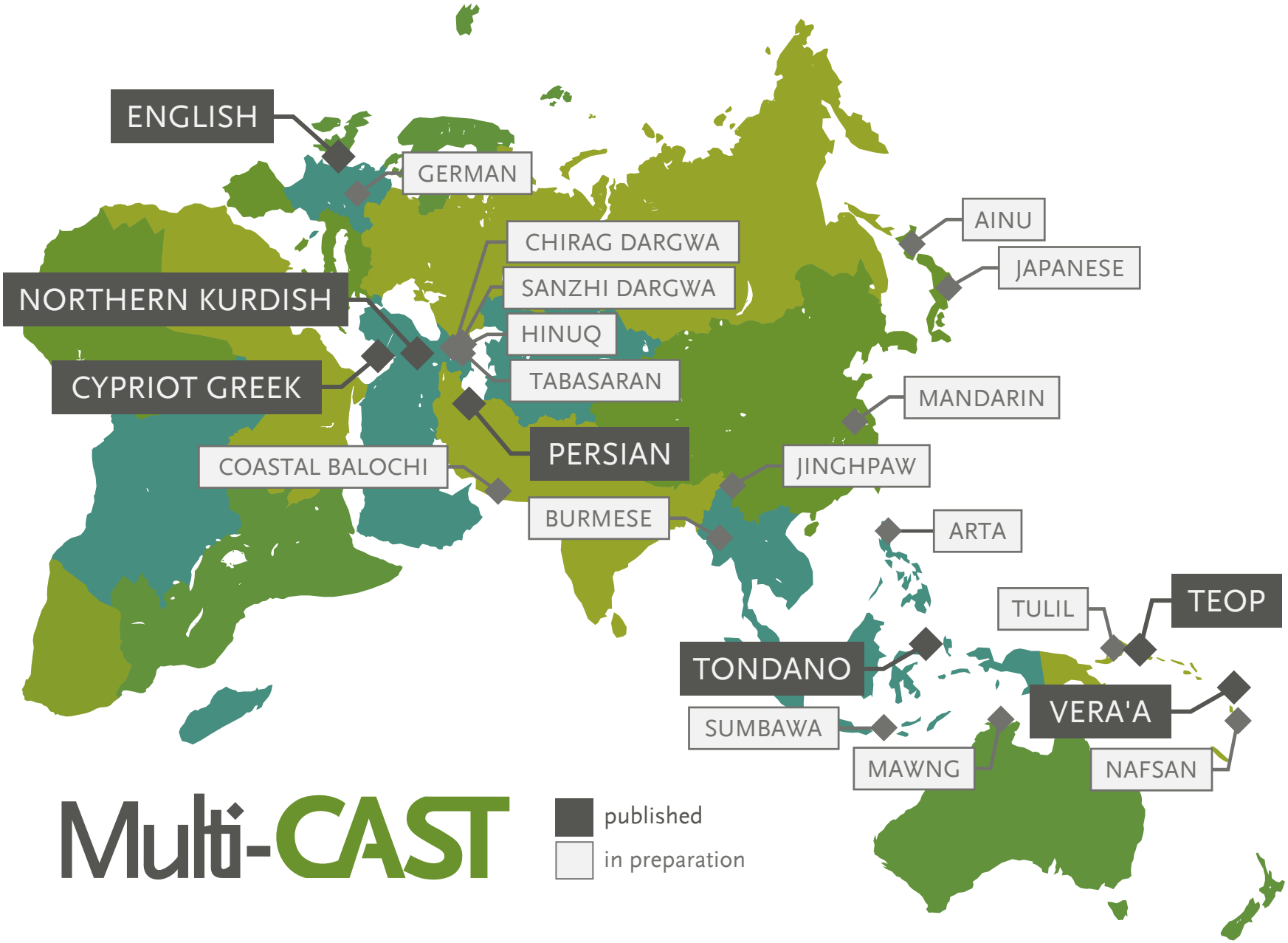
Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

https://lac2.uni-koeln.de/multicast/

— *normally at* —

https://lac.uni-koeln.de/multicast/


ENGLISH
GERMAN
NORTHERN KURDISH
CHIRAG DARGWA
SANZHI DARGWA
HINUQ
TABASARAN
CYPRIOT GREEK
PERSIAN
COASTAL BALOCHI
BURMESE
AINU
JAPANESE
MANDARIN
JINGHPAW
ARTA
TULIL
TEOP
TONDANO
VERA'A
SUMBAWA
MAWNG
NAFSAN
Multi-CAST
published
in preparation

References (1/3)

Andrews, Avery. **1985**. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description,* Volume 1, 62–154. 2nd edition. Cambridge: Cambridge University Press.

Andrews, Avery. **2007**. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description,* Volume 1, 132–223. 2nd edition. Cambridge: Cambridge University Press.

Chafe, Wallace. **1980**. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production.* Norwood, NJ: Ablex.

Du Bois, John. **1987**. The discourse basis of ergativity. *Language* 63(4). 805–855.

Du Bois, John. **2003a**. Argument structure: Grammar in use. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure: Grammar as architecture for function*, 11–60. Amsterdam: John Benjamins.

Du Bois, John. **2003b**. Discourse and grammar. In Tomasello, Michael (ed.), *The new psychology of language: Cognitive and function approaches to language structure*, Volume 2, 47–88. Mahwah, NJ: Erlbaum.

Du Bois, John. **2017**. "Ergativity and discourse in grammar." In Coon, Jessica & Massam, Diane & Travis Lisa D. (eds), *The Oxford handbook of ergativity,* 23–58. Oxford: Oxford University Press.

References (2/3)

Durie, Mark. **2003**. New light on information pressure: Information conduits, 'escape valves', and role alignment stretching. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure: Grammar as architecture for function*, 159–196. Amsterdam: John Benjamins.

Forker, Diana & **Schiborr**, Nils N. **In progress**. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*.

Haig, Geoffrey & **Schnell**, Stefan. **2014**. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators.* Version 7.0. (`https://lac2.uni-koeln.de/en/multicast/`)

Haig, Geoffrey & **Schnell**, Stefan. **2018**[2015]. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts.* (`https://lac.uni-koeln.de/en/multicast/`)

Haig, Geoffrey & **Thiele**, Hanna. **2015**. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (`https://lac.uni-koeln.de/en/multicast-northern-kurdish`)

Mosel, Ulrike & **Schnell**, Stefan. **2015**. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (`https://lac.uni-koeln.de/en/multicast-teop`)

References (3/3)

Riester, Arndt & **Baumann**, Stefan. **2017**. The RefLex scheme – Annotation guidelines. *SinSpec: Working papers of the SFB 732* (14). (http://elib.uni-stuttgart.de/handle/11682/9028)

Schiborr, Nils N. **2015**. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/en/multicast-english)

Schiborr, Nils N. & **Schnell**, Stefan & **Thiele**, Hanna. **2018**. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines*. Version 1.1. (http://bamling-research.de/multicast/web/data/1606/general/RefIND_guidelines_1.1.pdf)

Schnell, Stefan. **2015**. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/en/multicast-veraa)

Vollmer, Maria C. & **Hadjidas**, Harris. **2015**. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/en/multicast-cypriot-greek)