

Revisiting the Givenness Hierarchy

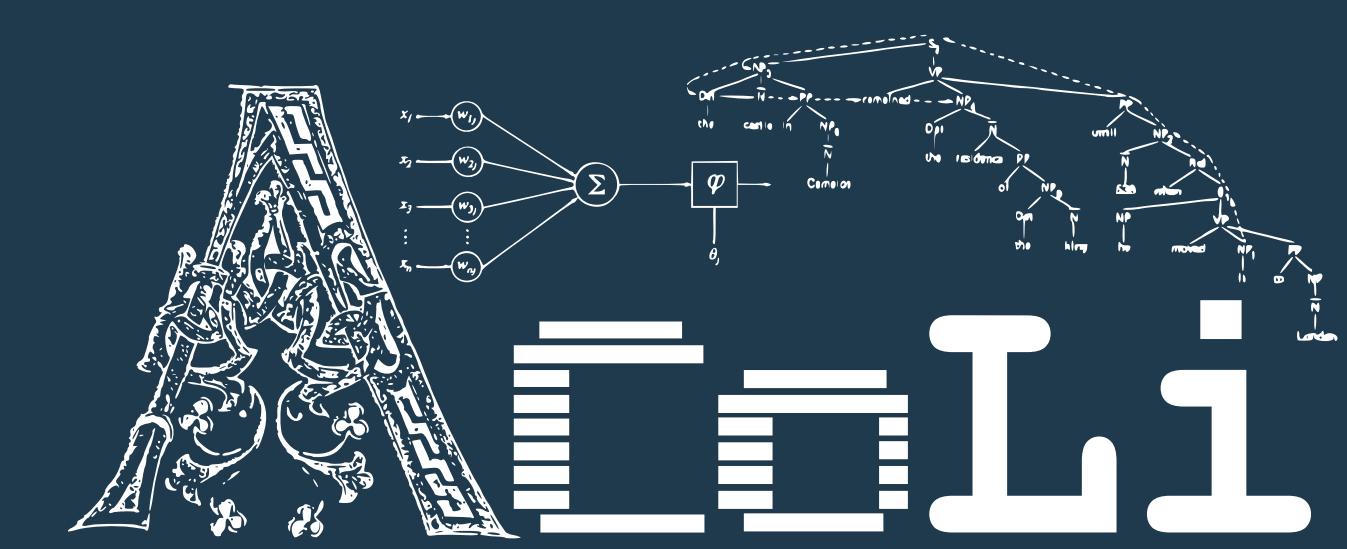
A Corpus-Based Evaluation



Universität
Augsburg
University

Christian Chiarcos

Applied Computational Linguistics (ACoLi)
University of Augsburg, Germany



Background and Motivation

Human communication is organized around **entities (discourse referents)**, their characteristics and their relations

For a speaker, effective communication usually requires

- to **keep track of entities** (supposed to be) known, identifiable or accessible to the hearer,
- to use communicative devices that allow their **unambiguous identification** by the hearer, and
- to use communicative devices that **indicate their relevance** with respect to the speaker's communicative goals

The **Givenness Hierarchy** (GH, Gundel et al. 1993) is one of several theoretical frameworks to account for this, which the following characteristics

- fine-grained predictions regarding the choice of referring expressions, esp. for demonstratives
- practical relevance (e.g., for language acquisition and human-robot interaction)*
- applied to a broad array of typologically diverse languages* * see references in paper
- comes with concise, cross-linguistically tested annotation guidelines ("coding protocol", 2006)

BUT

- almost no annotated data available
- some GH claims are controversial
 - personal pronoun > demonstrative pronoun ? (against Sgall et al. 1986)
 - demonstrative NP > definite NP ? (against Ariel 1990)
- previous studies were small-scale and usually did not involve statistic significance tests
 - results have not always been significant, especially for demonstratives

We suggest to

- replicate the original findings of Gundel et al. (1990, 1993)
- over **corpora with entity coreference corpora**, not direct annotation for GH
 - the coding protocol defines a decision tree for annotation, to a large extent based on (the form) of previous or subsequent mention
 - sufficient in size to expect significant results for less frequent types of referring expressions
 - available (now) for the original set of languages studied by Gundel et al. (1990, 1993)
 - Arabic, English, Chinese, Japanese, Korean, Russian and Spanish

Givenness Hierarchy (Gundel et al. 1993)

- hierarchy of „cognitive statuses“ ranked from highly given (**in focus**) to new, but identifiable in type (**type**)
- implicative hierarchy**
- higher status ⊑ lower statuses
 - e.g. referent r of „that man“ is **familiar**
 - r is also **unique, referential** and **type**
 - but not **in focus**, or **activated**
- cognitive statuses are lexicalized
 - referring expressions indicate the (expected) cognitive statuses as part of their lexical meaning
 - e.g. personal pronoun => **in focus**
- speakers deviate from lexicalized cognitive statuses to express implicit information
 - speakers can use a „lower“ expression to trigger **quantity implicatures**, e.g., when using a definite NP in place of a personal pronoun
- cross-linguistic annotation studies support the predicted correlations**
 - for 7 major languages in Gundel (1990/93), for other languages later on

e.g., English
(Gundel 1993)

seems to match
the predicted
correlations

re-assessment:
most GH-specific claims
are not statistically
significant
(for the original statistics)

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
it	214	1					215
HE		1					1
this			15				15
that		1	17				18
this N	1		11				12
that N			10	7			17
the N	30	95	47	108			280
indef. this N					1	1	
a N					41	55	96
total	246	150	54	108	42	55	655

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	
it	0.895++	-0.373++	-0.21++	-0.311++	-0.183+	-0.212++	1.0 < x < 0.5 ***
HE	-0.03 n/a	0.072 n/a	-0.012 n/a	-0.017 n/a	-0.01 n/a	-0.012 n/a	0.5 < x > 0.1 ***
this	-0.119 n.s.	0.281 n/a	-0.046 n/a	-0.068 n/a	-0.04 n/a	-0.046 n/a	0.1 > x > 0.0 ***
that	-0.111+	0.286 n/a	-0.05 n/a	-0.075 n/a	-0.044 n/a	-0.051 n/a	n.s. > 0.1 ***
this N	-0.062(+)	0.224 n/a	-0.041 n/a	-0.061 n/a	-0.036 n/a	-0.041 n/a	0.2 > x > -0.5 ***
that N	-0.127(+)	0.14 n/a	0.195 n/a	-0.073 n/a	-0.043 n/a	-0.049 n/a	-0.5 > x > -1.0 ***
the N	-0.479++	0.227++	0.268++	0.514++	-0.226++	-0.262++	
indef. this N	-0.03 n/a	-0.021 n/a	-0.012 n/a	-0.017 n/a	0.149 n/a	-0.012 n/a	
a N	-0.479++	0.227++	0.268++	0.514++	-0.226++	-0.262++	

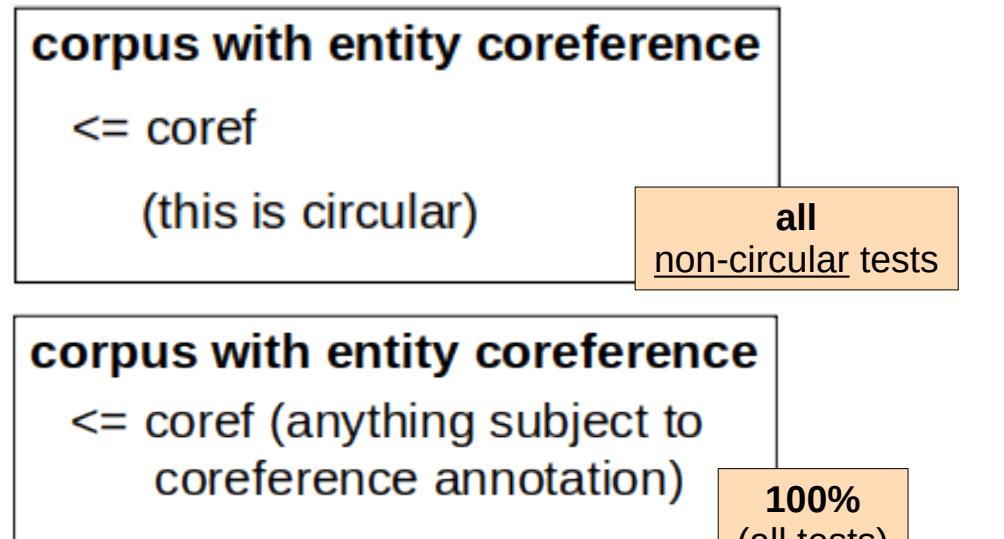
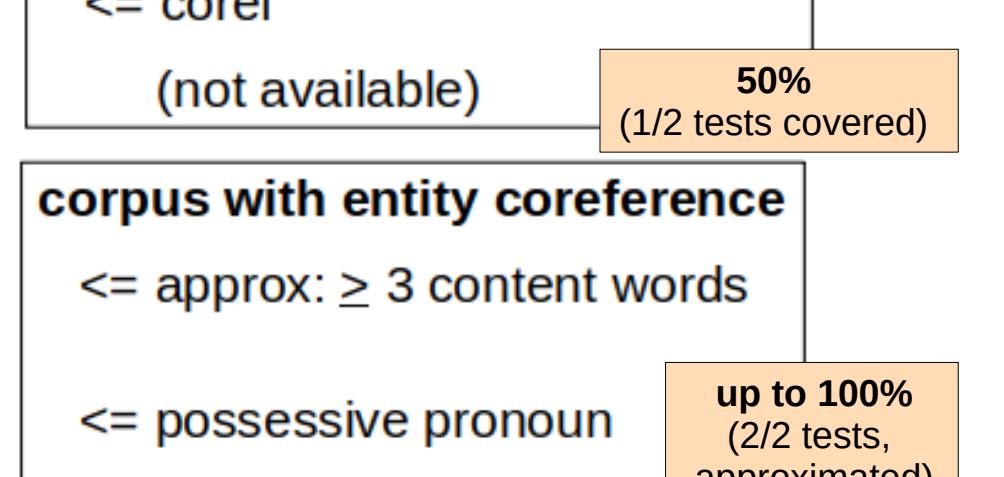
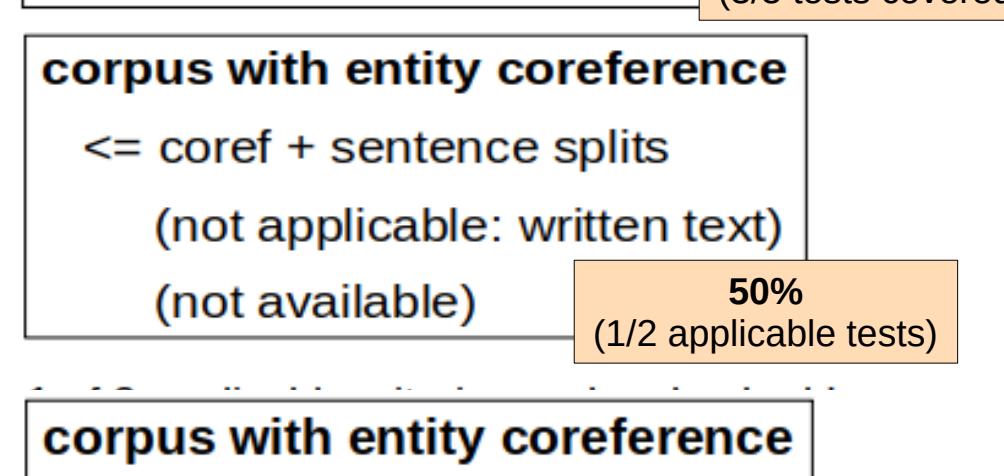
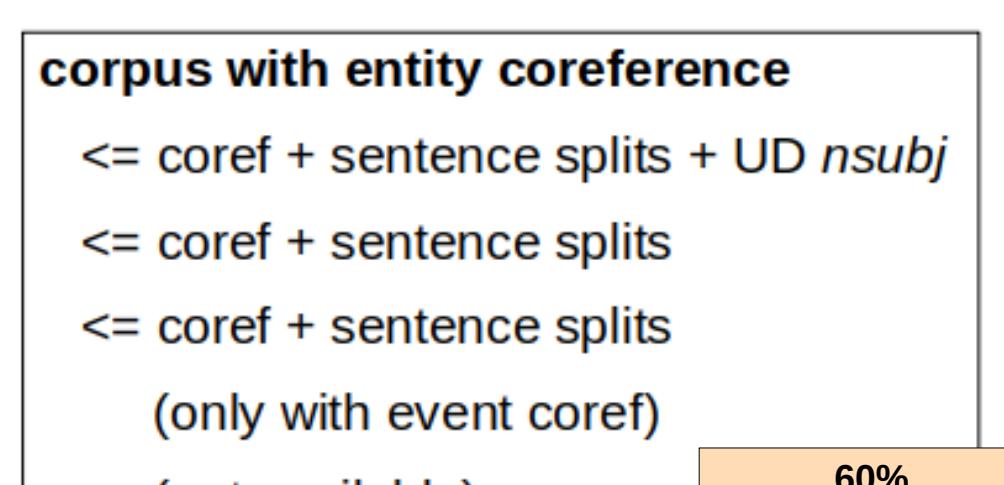
n.a. * not applicable
n.s. not significant
(*) marginal, p<0.05
+ significant, p<0.01
++ highly significant, p<0.001

Bootstrapping Givenness from Entity Coreference

Coding protocol (2006):

- for every referring expression, **check the statuses from highest to lowest**, annotate the highest possible status
- for every status, a number of test formulate criteria sufficient for annotation

- annotate **in focus** if
 - r is subject of the preceding utterance
 - r mentioned earlier in same utterance
 - r mentioned in both of the two previous utterances
 - r is the event of the preceding utterance
 - r is an inferred discourse topic
- annotate **activated** (if not **in focus** and)
 - r is mentioned in the two previous utterances
 - r evoked by gesture or gaze
 - r is an associated proposition or speech act
- annotate **familiar** (if not **activated** or **in focus** and):
 - r previously mentioned
 - r known from shared background
- annotate **unique** (if not **familiar**, etc.):
 - expression contains sufficient lexical material to create a unique referent
 - r linked via lexical association to activated referent
- annotate **referential** (if not **unique**, etc.):
 - r mentioned later in discourse
 - r linguistically marked for discourse prominence
- annotate **type** (if not **referential**, etc.):
 - expression encodes interpretable conceptual content



if coreference annotation fails to capture a criterion, the bootstrapped status may be lower than the actual status (but never higher) because the hierarchy is implicative, this is imprecise, but not incorrect

Empirical Evaluation

10 corpora, 7 languages, 4 coref formats, UD syntax (as provided / spaCy) up to 11 types of referring expressions* per language

* exactly those defined by Gundel et al. (1990/1993)

	OntoNotes	LitBank	GUM	AnCora	NTC	KoCoNovel	ECMT	RuCor
version	5.0	CU 1.3	CU 1.3	CU 1.3	1.5	—	CU 1.3	CU 1.3
language	ar / en / zh	en	en	es	ja	ko	ko	ru
modality	written	written	written/spoken	written	written	written	written	written
genre	news, web, lit	literature	diverse	news	1,000	165	439	145
tokens (K)	325 / 1,750 / 235	190	170	429	1,000	165	439	145
pron	3066	258	323	117	4720	—	—	—
dem.prox	105	38	72	39	416	—	—	—
dem.dist	117	42	1	35	34	403	—	—
dem.prox N	155	156	130	109	216	982	—	—
dem.dist N	101	19	54	34	90	374	—	—
the N	1046	1082	1169	1021	2921	8149	—	—
a N	691	123	731	416	1628	3719	—	—
total	5281	2617	1718	2092	2010	5045	18763	18763

Evaluation for English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	
pron	0.475++	0.103+++	0.074++	-0.073+++	-0.319++	1.0 < x < 0.5 ***	
dem.prox	-0.105 n.s.	0.109 n.s.	0.053	-0.053	0.032+++	0.5 < x > 0.1 ***	
dem.dist	-0.127 n.s.	0.125+++	0.077 n.s.	-0.051+++	-0.01 n.s.	0.1 > x > 0.0 ***	
dem.prox N	-0.065++	0.056+++	0.055+++	0.016(+)	0.003 n.s.	0.026+++	n.s. > 0.1 ***
dem.dist N	-0.004 n.s.	0.026++	-0.02++	0.015(+)	-0.007 n.s.	-0.009 n.s.	0.2 > x > -0.5 ***
the N	-0.238+++	-0.074++	0.125+++	0.089+++	0.051+++	0.177++	0.5 < x > 0.1 ***
a N	-0.106+++	-0.15++	-0.101+++	0.134+++	0.008 n.s.	0.189++	

sample results (GUM corpus, Zeldes 2017)

aggregate results, 4 corpora: we count significant (p < 0.01) positive and negative correlations

GH predictions largely confirmed, possible exception that N (dem.dist N)

Aggregate Results, 6 non-English Corpora

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	es, zh	ja,zh	es,ar	ar	es,zh, zh, ar, ru, es, zh, ru, zh, ru, zh, ru, zh, ru, zh, ru	