Revisiting the Givenness Hierarchy. A Corpus-Based Evaluation

Christian Chiarcos
Applied Computational Linguistics (ACoLi)
University of Augsburg, Germany

Abstract

The Givenness Hierarchy (GH) models how speakers signal cognitive statuses of referents in discourse, playing a key role in computational models of situated communication and in applied linguistics. We present an empirical method to evaluate the Givenness Hierarchy using large corpora with coreference annotations. Our findings confirm predicted associations between cognitive statuses and referring expressions across multiple languages, while also highlighting limitations, notably difficulties to approximate the cognitive status UNIQUE and to account for demonstrative noun phrases. Additionally, we demonstrate how coreference data can be used to bootstrap GH annotations, facilitating automatic labelling of cognitive statuses and advancing discourse-aware NLP. Finally, we provide conversion scripts to transform Japanese and Korean corpora into CorefUD-compatible formats, supporting broader multilingual research despite current annotation and licensing constraints. Our work bridges theoretical linguistics and practical computational methods, offering a scalable framework to study givenness across diverse languages.

1 Background and Motivation

Effective reference – whether by humans or dialogue systems – requires contextually appropriate expressions. As illustrated by variations in translation, language offers multiple ways to express the same referent: While Challoner (1749-1752 (revision) translated 1Ki, 11,28 with pronoun and elision (*Solomon seeing him (...), Ø made him chief ...*), Darby (1890) used a definite NP and a pronoun (*Solomon saw the young man (...), and he made him ruler ...*). To account for the functional dimension of this flexibility, various theories of information status (Prince, 1981; Givón, 1983; Ariel, 1990; Chafe, 1994) posit hierarchies or scales mapping referential forms to degrees of 'salience', 'accessibility', or 'givenness', but while they agree on broad trends – pronouns denote high degrees of givenness, full NPs low – they differ in terminology and granularity. From the set of prominent theories, we adopt the Givenness Hierarchy (GH) by Gundel et al. (1993) for three reasons: (1) It is a relatively detailed theory in that it accounts not only for pronominal, nominal, definite and indefinite forms, but also for different qualities of demonstrative pronouns and demonstrative NPs, as well as for possible deviations from the expected encoding of the statuses it proposes; (2) unlike any other of the aforementioned theories, it comes with explicit, and practical annotation guidelines;[1] and (3) the theory and its annotation guidelines have been applied to a considerable number of typologically diverse languages.[2] Moreover, the GH is while widely cited in technical contexts (Han et al., 2022; Pal et al., 2021; Spevak et al., 2022; Higger and Williams, 2024; Daigler et al., 2024),[3] as well as in applied linguistics (Gundel and John-

[1] There are other annotation guidelines for information status, e.g., Nissim et al. (2004); Ritz et al. (2008); Baumann and Riester (2013); Dyer et al. (2024), but these aim to generalize over multiple theories and are thus not directly comparable. In particular, they cannot be directly used to evaluate specific claims of Gundel et al.'s theory if the underlying theories did not share the same predictions, esp., regarding the use of demonstratives.

[2] Aside from the major languages considered here and by Gundel et al. (1990, 1993), this also includes Breton (Hedberg and Schapansky, 1996), Yapese (Ballantyne, 2004), Kumyk (Humnick, 2005), Irish (Mulkern, 2008), Kaqchikel Maya (Hedberg, 2010), Eegimaa, Ojibwe (Gundel et al., 2010), Farsi (Khormaee and Skrouchi, 2015), Luo (Omondi, 2016) and American Sign Language (Swabey, 2011), among others.

[3] In the era of LLMs, many of the challenges that theories of information status such as the Givenness Hierarchy account for – reordering constituents, anaphor resolution and generation, prediction and interpretation of non-canonical structures, lexicalizing frames, and handling grammatical voice – may be solved to some extent in practice, but only for major languages and but without any insight into the underlying processes and their actual effects on the interlocutors, thus not directly applicable for low-resource languages or in controlled and vulnerable settings such as in human-robot dialog.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 24–41
Suzhou, China, November 9, 2025.

son, 2013; Kim, 2016; Velnic, 2018; Krüger, 2018). At the same time, however, its empirical basis remains limited. Early support came from elicitation experiments (Gundel et al., 1990, 1993), but few annotated corpora are publicly available, the limiting replicability and application. Here, we provide a technical operationalization of the Givenness Hierarchy on the basis of existing corpora with coreference annotation for all languages originally considered by Gundel et al. (1990, 1993) to motivate and develop their theory (English, Arabic, Chinese, Japanese, Korean, Russian, and Spanish), with the goal of to evaluate the theory, to test and confirm its predictions, and to develop a method for bootstrapping language-specific givenness hierarchies for other languages from existing coreference annotation.

Gundel et al. (1993) postulate a hierarchy of six cognitive statuses and their alignment with prototypical expressions. Below, these ranked from most to least accessible

1. **in focus**: referent is the current focus of attention and highly prominent in the local context ($\sim$ *he*, *she*).
2. **activated**: referent is present in the local context ($\sim$ *this/that*, *this man*).
3. **familiar**: referent is known to both speaker and hearer from prior discourse ($\sim$ *that man*).
4. **unique**: the referent is uniquely identifiable to hearer and speaker ($\sim$ *the man*).
5. **referential**: the speaker refers to a specific but possibly unknown entity ($\sim$ *this guy*).
6. **type** (type identifiable): hearer can identify the category of a referent ($\sim$ *a man*).

Unlike other theories, this is an implicative hierarchy, i.e., a status higher in the hierarchy entails all lower statuses, so, it can be referred to with their forms – and speakers may use such deviations to convey implicatures (e.g., using a definite NP for an in-focus entity to emphasize contrast). Gundel et al. (1990,1993) provide predictions for English, Arabic, Chinese, Japanese, Korean, Russian, and Spanish and evaluate these in elicitation experiments. However, no annotations seem to be publicly available, and the numbers they (and many later papers) report do not always reach the levels of statistical significance. Table 1 replicates their English data and adds Pearson correlation and binary χ^2 significance scores for each pairing of referring expressions and statuses, and this reassessment confirms key GH assumptions: significant positive correlation between **in focus** and pronouns, **unique** and definite NPs, **type** and indefinite NPs. It also shows negative correlations where expected. However, data on demonstratives and emphatic pronouns remain sparse, indicating a potential weak spot in theory – but, ironically, these are the very predictions that distinguish GH from competing theories: While the general pattern of pronoun $>$ definite NP $>$ indefinites is generally accepted, its fine-grained distinctions, especially regarding demonstratives, remain controversial. Siddharthan et al. (2011) argue that GH conflates dimensions, and psycholinguistic experiments such as Xu and Xiang (2021) failed to confirm some predicted effects. The status of demonstratives is particularly contested: In direct opposition to Gundel et al. (1993), Sgall et al. (1986) claimed that demonstrative pronouns rank higher than personal pronouns, and Ariel (1991) saw demonstrative NPs as lower than definites.

To address data sparsity and controversies – but also potential biases of annotators who are aware that certain forms indicate certain categories –, we propose a new approach: extrapolating cognitive status from existing corpora with coreference annotation, which substantially exceed the traditional elicitation experiments in scale. In the last years, this approach has become feasible due to the increased availability of corpora with coreference annotation, covering now all original GH languages. In comparison with earlier elicitation methods, these offer higher coverage – hundreds or thousands of tokens per referential form – and reduce circularity risks. We replicate the original GH setup by focusing on the same referential expressions,[4] to the extent they are annotated.[5] By grounding the evaluation of the Givenness Hierarchy in independently created coreference corpora, we aim to reassess its predictions and offer a scalable, reproducible methodology to support or revise its theoretical foundations. We are specifically interested in debated GH claims that are not suffi-

[4]This is particularly important for calculating totals. In particular, we do not evaluate against *all* referring expressions, but only against (the total of) those considered by Gundel et al., so that certain categories, e.g., first- and second-person anaphora, pronominal adverbs, quantified NPs and proper names are excluded.

[5]Some corpora have an annotation bias, e.g., we have no annotation of zero anaphora (Ø) for Korean and Russian, and some corpora, in particular, OntoNotes (Pradhan and Xue, 2009), only provide annotations for specific referents, effectively neglecting the **type** category. No corpus we worked with has annotations of event anaphora.

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
it	0,895++	-0,373++	-0,21++	-0,311++	-0,183+	-0,212++
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
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
that	-0,111+	0,286 n/a	-0,05 n/a	-0,075 n/a	-0,044 n/a	-0,051 n/a
this N	-0,082(+)	0,224 n/a	-0,041 n/a	-0,061 n/a	-0,036 n/a	-0,041 n/a
that N	-0,127(+)	0,14 n/a	0,195 n/a	-0,073 n/a	-0,043 n/a	-0,049 n/a
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++

1.0 > r > 0.5 +/++
0.5 > r > 0.1 +/++
0.1 > r > 0 +/++
n.s / (+) / n.a
0 > r > -0.1 +/++
-0.1 > r > -0.5 +/++
-0.5 > r > -1.0 +/++

n/a	χ^2 not applicable
n.s.	not significant
(+)	marginal, p<=.05
+	significant, p<=.01
++	highly significant, p<=.001

Table 1: Cognitive statuses and referential expressions in English (Gundel et al., 1993), with correlation and χ^2 significance, Gundel et al.'s absolute numbers are provided in Tab. 7

ciently substantiated by earlier empirical analyses, in particular, the relative givenness of demonstratives in comparison to pronouns and definite NPs.

2 Experimental Setup

We provide an operationalization of the Givenness Hierarchy on the basis of the original annotation guidelines (Gundel et al., 2006). These instruct annotators to assign the highest applicable cognitive category according to the following overview. In addition to the original instructions, we indicate whether a criterion is directly implementable (✓), can be heuristically approximated ((✓)), or not operationalizable (?). We also provide approximation criteria, but only operate with those that do not introduce dependencies from surface forms. These are marked by (✓*).

Given a referring expression e and referent r:

- annotate **in focus** if
  - ✓ r is subject of the preceding utterance (∼ `nsubj`)
  - ✓ r mentioned earlier in same utterance
  - ✓ r mentioned in both of the two previous utterances
  - (✓) r is the event of the preceding utterance (∼ neuter weak pronouns without antecedent)
  - ? r is a discourse topic inferred but not overtly mentioned
- **activated** (if not **in focus** and):
  - ✓ r is mentioned in one of the two previous utterances
  - (✓*) r evoked by gesture or gaze (n/a, we operate with written text)
  - ? r is an associated proposition or speech act
- **familiar** (if not **activated** or **in focus** and):
  - ✓ r previously mentioned
  - ? r known from shared background
- **unique** (if not **familiar**, etc):
  - (✓*) e contains sufficient lexical material to create a unique referent (∼ by the use of more than 3 content words)
  - (✓) r linked via lexical association to activated referent (∼ possessive pronouns)
- **referential** (if not **unique**, etc.):
  - ✓ r mentioned later in discourse
  - (✓) r linguistically marked for discourse prominence
- **type** (if not **referential**, etc.):
  - (✓*) e encodes interpretable conceptual content (∼ anything subject to coreference annotation)

Out of 15 criteria, 6 are directly implemented, 6 can be approximated, and 3 not covered. By including surface criteria, we can cover up to 80% of the original protocol – although, here, we decided to remain agnostic about surface forms to avoid circular reasoning and operate only with (approxiations for) 9 criteria (60%). While this introduces some noise, we assume it will not preclude meaningful generalizations if statistically significant patterns emerge. Additional design decisions include equating 'utterance' with sentences (based on provided parse or produced by a parser), and the definition of 'mentioned in' as 'having an anaphor/antecedent annotated in'. For pro-drop languages without Ø annotation (Russian, Korean), this leads to an underrepresentation of **in focus** and **activated**. As none of our corpora systematically annotates event anaphora; such cases may be wrongly treated as discourse-new.

Aside from Gundel et al. (2006), there are alternative GH operationalizations that reflect language-specific needs or annotation trade-offs,[6] but mostly

[6]Alternative operationalizations of the GH include Henschel et al. (2000), who redefine **in focus** as subject of last

represent simplifications. We operate with Gundel et al. (2006) to order to follow Gundel et al.'s original six-way distinction.

3 Empirical Evaluation

Based on corpora with coreference annotation, and using the heuristics described above, we compute Pearson's r and assess correlation significance using the χ^2 test for each pairing of cognitive status and the type of referring expression. This section provides aggregate results, with full data in the Appendix.

Table 2 gives an overview over the corpora considered, using OntoNotes v5.0 (Pradhan and Xue, 2009), NTC 1.5 (Iida et al., 2017), KoCoNovel (Kim et al., 2024), and the CorefUD 1.3 (Nedoluzhko et al., 2022) editions of AnCora (Recasens and Martí, 2010), ECMT, GUM (Zeldes, 2017), LitBank (Dyer et al., 2024), ParCorFull (Lapshinova-Koltunski et al., 2018), and RuCor (Ju et al., 2014).[7] These corpora vary in scope, genre, and annotation practices. Not all provide full coreference coverage – OntoNotes only annotates specific, referential entities, while KoCoNovel restricts to protagonists, omitting inanimates. Event anaphora are generally excluded. Because of the resulting noise, it is thus important to interpret them in conjunction with data from Gundel et al. (1990) and later papers that is smaller in scale but produced under more controlled conditions. As for CorefUD corpora, we use the existing UD annotations for classifying referring expressions and extracting grammatical features (esp. `nsubj`), for other corpora, we operate with automated parses obtained from spaCy, resp., for Arabic, spaCy-udpipe (i.e., UDpipe 2.5). We operate with sentence boundaries as provided and use those predicted by the parsers where these are lacking.

We distinguish the following types of referring expressions (cf. Tab. 6 for other languages) :

Ø zero anaphor (if annotated)
pron third-person pronoun (e.g., *it*)
dem.prox proximal dem. pronoun (e.g., *this*)
dem.med medial dem. pronoun (e.g., Span. *ese*)
dem.dist distal dem. pronoun (e.g., *that*)
dem... N demonstrative NP (e.g., *this house*)
def N definite NP (e.g., *the house*)
Ø N bare NP (e.g., Russian *dom*)
ind N indefinite NP (e.g., *a house*)

These categories are language-specific and functionally not always equivalent. For instance, medial demonstratives exist only in Spanish, Japanese, and Korean. Zero anaphors vary in distribution and constraints; e.g., Spanish allows them for subjects, only, but Japanese also for objects. The category `Ø N` also has different functions languages, depending on (the lack of) a grammatical opposition with indefinite or definite NPs.

The Givenness Hierarchy postulates three principles to account for deviations from the expected associations between forms and cognitive statuses: (1) all statuses can be expressed with forms for lower statuses (implicative hierarchy), (2) deviations are used to trigger Gricean quantity implicatures (and thus, rarer than non-deviations), and (3) these deviations are monodirectional (downward only). Statistically, we thus expect positive correlations between statuses and their prototypical form, absence of low statuses encoded with higher-status forms, and noise from our heuristic-based status approximations, with a possible overrepresentation of **type** (and, possibly, **referential**) for actual cases of event anaphora. Tables 3 and 4 summarize our results with aggregate correlation data for English and other languages, respectively. Overall, the reported correlations are statistically significant, but low, at times, reflecting both imperfections in the annotation-based cognitive status approximations and noise in the data.

In all English corpora (Tab. 3), `pron` correlates positively with **in focus** and **activated**, and negatively with lower statuses, consistent with Gundel et al. Demonstrative pronouns correlate positively with **activated** and negatively with **in focus**, supporting their distinct status from personal pronouns. Unexpected positive correlations with **referential** and **type** (in OntoNotes) may result from event anaphora.

Proximal demonstrative NPs exhibit negative

utterance plus *all* discourse-old entities, with an accompanying salience ranking based on Grosz et al. (1995). Although Gundel (1998) also explore this link, they reject scalar models, preferring categorical distinctions. Henschel et al.'s simplification thus collapses several GH categories into the equivalent of **familiar**. Another simplification is Traugott et al.'s (2008) use of only three GH statuses (**familiar**, **unique**, **referential**) for Old English. For Spanish, (Blackwell and Quesada, 2012) merged **referential** and **type**, and further distinguish between recoverable and non-recoverable **activated** referents. However, it is unclear how to technically operationalize their notion of recoverability on the basis of coreference annotations.

[7] Note that for Arabic, OntoNotes reflects Modern Standard Arabic, which may differ from the spoken varieties used by (Gundel et al., 1990, undocumented variety) and (Gundel et al., 2010, Tunesian Arabic).

	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	news	literature	news	diverse
tokens (K)	325 / 1,750 / 235	190	170	429	1,000	165	439	145

Table 2: Coreference corpora considered (CU = CorefUD)

correlation with **in focus** and the expected positive correlation with **activated**, but also with **familiar**, contrary to Gundel et al. This suggests a possible reclassification aligning them with **familiar**. Isolated positive correlations with **referential** (in OntoNotes) are in line with predictions for indefinite *this*, but may be due to the incomplete nature of coreference annotation. Distal demonstrative NPs correlate positively with **activated** and **unique** in OntoNotes, and with **referential** and **type** in LitBank. Their inconsistent behavior suggests Gundel et al.'s hierarchy may not fully explain their use. Other functions appear to be likely, e.g., the use of distal demonstratives in comparisons with referents referred to with proximal demonstratives. Overall, proximal forms are more frequent (e.g., OntoNotes: 3743 vs. 1904; GUM: 982 vs. 374), possibly due to their broader use contexts, and these (but only these), seem to adhere to the expected distribution.

Definite NPs correlate strongly with **familiar** and **unique**. Correlations with lower statuses may reflect limitations of the approximation of **unique** by lexical richness. Still, they tend to encode lower statuses than demonstratives. As for indefinite NPs, these are negatively correlated with previous mention (**familiar** or higher) and positively correlated with the lack thereof (**unique** or lower). Again, the differentiation between **unique** and lower statuses may be insufficient to delineate the narrower scope of indefinite NPs.

As for the cognitive statuses themselves, we see good evidence for **in focus** (positively correlated with the use of third-person pronouns, and only these) and **activated**) (positively correlated with demonstrative pronouns. and only these), as well as for the differentiation between **familiar** (negatively correlated with indefinite NPs) and lower statuses (positively correlated with indefinite NPs). The **unique** status can probably not be approximated from coreference annotations that would be sufficiently reliable to be used in a meaningful way in this evaluation. This is different for **referential**, which can be easily identified from coreference annotation. Yet, in the corpora of (mostly) written language considered here, there are no referring expressions that seem to require a differentiation between **referential** and **type**.

For other languages (Tab. 4), pronouns and Ø correlate with **in focus**, as predicted. The positive correlations between Ø and **activated** in Japanese and Chinese may be due to the fact that these languages do not limit zero anaphors to subject antecedents and can have more than one Ø as argument. The positive correlations between third-person pronouns and ACTIVATED for Russian, Japanese and Chinese may reflect that in these languages, pronouns can be more easily replaced by zero anaphors, so that overt pronouns are more likely to take on characteristics of stressed pronouns ... that Gundel et al. (1993) associate with **activated**. Demonstrative pronouns correlate with **activated** but not with **in focus** or with lower statuses, with the possible exception of Korean. Some **type** correlations may stem from exophoric or event-based reference, especially in dialogue.

Similar to English, demonstrative NPs appear heterogeneous and hard to interpret. Their negative correlation with **in focus** is in line with Gundel et al., and they seem to be associated with cognitive statuses at the same level or below demonstrative pronouns. Noteworthy is the systematic association between proximal demonstrative NPs and **activated**, which is predicted by Gundel et al. Indeed, English seems to be exceptional in this regard in extending the scope of *this*-NPs to **familiar**. For medial and distal demonstrative NPs, there seems to be no clear positive correlation with any cognitive status. This, again, indicates that the functions of demonstrative NPs (other than proximal demonstrative NPs) may involve other functions than givenness marking. Definite NPs correlate with previous mention (i.e., **familiar**, for Spanish), but aside from their negative correlation with **in focus**, they can be used for any status at the level of **unique** or higher. That we also find correlations with **referential** and **type** may be due to the insufficiencies of our approximation of **unique**, as this differentiation, indeed, was already statistically

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	4:0	2:1	0:4	0:3	0:4	0:2
dem.prox	0:1	2:0		0:2	1:1	0:1
dem.dist	0:1	2:0	0:1	0:2	0:1	1:1
dem.prox N	0:4	3:0	3:0	0:1	1:1	0:1
dem.dist N	0:2	2:0	0:1	1:0		
the N	0:4	1:2	4:0	3:0	4:0	2:0
a N	0:4	0:3	0:3	3:0	2:0	2:0

4:0
3:0
2:0, 3:1
1:0, 2:1
1:1
0:1, 1:2
0:2, 1:3
0:3
0:4

Table 3: Aggregate correlations for four English corpora; green indicates significant positive correlations, red negative ($p \leq 0.01$).

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	es,ja, zh,ar	ja,zh es,ar	ar	es,ja, zh,ar	es,ja, zh,ar	es,ja
pron	ru,es,ja, zh,ar	ru,ja,zh es,ar	ru,es,zh, ar (ja)	ru,es, zh,ar	ru,es, zh,ar	
dem.prox	ko1 ja	es,ja,zh	ko2,zh es,ko1	es,zh	es,ko1, zh	ko
dem.med	ko ja	es,ko,ja	ko1	es, ko2 (ja)		es,ko ko
dem.dist	ko2 zh	zh				es
dem.prox N	ko2	es, ja,zh	ru,es, ko1,ja,zh	zh	ko1 (ja)	es zh ru,es, ko1,zh
dem.med N	ko1,ja	ko1	(ja)	es,ko1	es	es,ko1 ko1
dem.dist N	zh	zh	zh			ru
the N	es,ar	es ar	es	es ar	es ar	es
N	ru,ko, ja,zh	ru,ko, ja,zh	ru,ko2, zh (ja)	ru, ja,zh ko1	ru,ko, ja,zh	ko, ja
a N	es,zh, ar	es,zh, ar	es,zh,ar	zh,ar	es,zh,ar	

ar	Arabic (OntoNotes)
es	Spanish (AnCora)
ja	Japanese (NTC)
ko1	Korean (ECMT only)
ko2	Korean (KoCoNovel only)
ko	Korean (ECMT=KoCoNovel)
ru	Russian (RuCor)
zh	Chinese (OntoNotes)

4:0, 5:0
3:0
2:0, 3:1
1:0, 2:1, 3:2
1:1, 2:2
0:1, 1:2, 2:3
0:2, 1:3
0:3
0:4, 0:5

Table 4: Aggregate correlations for Arabic, Chinese, Japanese, Korean, Russian, and Spanish. Significant correlations ($p \leq 0.01$): positive on the left, negative on the right of each status cell.

significant in Gundel et al.'s original data. As for Ø NPs, these are negatively correlated with **in focus** and **activated**, and positively with **familiar** and lower statuses. This underlines the relevance of the differentiation between **activated** and **familiar**, but also that Ø NPs cover functions otherwise adopted by definite and indefinite NPs and are thus applicable to *any* cognitive status, if the conditions are met. The correlations of indefinite NPs underline the importance to differentiating previous mention (**familiar**) and the lack thereof (**unique** or lower).

As for the evaluation of cognitive statuses, the distribution differences support the differentiation of **in focus** (for Ø and pron), **activated** (for stressed pronouns and demonstrative pronouns), **familiar** (for Ø NPs in languages without grammaticalized determiners) and statuses lower than **familiar** (for indefinite NPs). Again, **unique** cannot be reliably differentiated from lower statuses with the heuristics adopted here, but whereas **referential** can, it does not seem to be necessary to account for any of the major classes of referring expressions.

Similarly as for the case of English, we thus find that pronouns mark higher givenness than demonstratives, demonstrative pronouns and *proximal* demonstrative NPs rank above definites and bare NPs. Medial/distal demonstrative NPs resist straightforward classification and may involve other pragmatic functions beyond those captured by the Givenness Hierarchy.

4 Consolidation, Inference and Revision

We would like to combine our findings with those of Gundel et al. (1993) – who report statistically significant differences between **unique** and **referential** when accounting for definite NPs –, and

overall suggest to reconsider the status of **referential**. While absent from the original proposal (Gundel et al., 1990), this was introduced by Gundel et al. (1993) specifically to account for indefinite *this* in English, but have failed to demonstrate its (statistical) significance, and neither do our correlations call for a differentiation between **referential** and **type identifiable**.

For this reduced hierarchy of *five* cognitive statuses, we find robust evidence both in our data and Gundel et al.'s: correlations between **in focus** and pronouns/Ø; between **activated** and demonstratives/proximal demonstrative NPs; between **familiar** and the avoidance of indefinite NPs; between **unique** and definite NPs (per Gundel et al.); and between lower statuses and indefinite NPs. However, medial and distal demonstrative NPs remain difficult to classify: neither our data nor Gundel et al.'s show significant correlations, suggesting these are sensitive to factors other than givenness.

Further, the approach enables bootstrapping Givenness Hierarchies in other languages or for phenomena that Gundel et al. did not originally consider. For instance, Mulkern (2011) investigates the referential properties of proper names, distinguishing between full names (e.g., given name plus surname) and single names (e.g., family name alone or nicknames). She suggests that full names align with **unique** and single names with at least **familiar**.

Using the correlation analysis described above, we can now verify these claims in an empirical quantitative manner. We approximate the notion of single name by single token proper names, full names by multi-token proper names and perform the evaluation against the OntoNotes corpus, as it is by far the largest corpus in our sample. The results (Tab. 5) confirm, indeed, that proper names are associated with the middle segment of the Givenness Hierarchy (thus negative correlation with **in focus**), and moreover, that short names, or, at least, single-token names, tend to be associated with higher cognitive statuses than full, resp., multi-token names, as these differ in their correlation with **activated**. As mentioned before, the approximation of **unique** from coreference annotations is insufficient, but on conceptual grounds – as pointed out by Mulkern –, any element that a hearer can recognize as a proper name is by definition **unique**. An interesting observation is, however, that the cognitive statuses that full and short names seem to be associated with are not **familiar** and **unique**, as postulated by Mulkern, but, rather, **activated** and **unique**. However, this may be an artifact of the heuristic approximation of single names by single tokens and full names by multi-word expressions, as a considerable number of single tokens will indeed just represent the complete name of, say, a country, and these might behave differently from person names.

We can confirm the general pattern of short names associated with higher givenness and long names associated with lower givenness, but it also seems that further differentiations within the larger group of proper nouns are likely, with their own alignments with cognitive statuses, and that these align with, but complement the distinction of different kinds of proper names studied by Mulkern. With this in mind, future research may now explore more fine-grained distinctions of referring expressions in an empirical fashion, and, potentially, revise the Givenness Hierarchy.

For future studies of the Givenness Hierarchy with coreference-annotated corpora, we suggest a to operate with a simplified model where the current **referential** category is abandoned. However, unlike (Traugott and Pintzuk, 2008), (Blackwell and Quesada, 2012), and (Abisambra Miccheli and Quesada, 2023), we do not suggest to merge it with **type**, but, instead with **unique**, as, according to Gundel et al. (1993), it is a superset of **unique**, and it provides clear, verifiable criteria for its distinction from **type**: **referential** can be inferred from subsequent anaphora. To avoid ambiguity, we propose renaming this unified category to **XREF** (extended referential). While our data could not distinguish **referential** from **type** conclusively, future research may uncover such distinctions for **XREF**. With truly **unique** referents accumulating in this category, we would expect that some of the effects (Gundel et al., 1993) and later studies found for **unique** are detectable in this broader category.

5 Results and Perspectives

We describe the empirical verification of the Givenness Hierarchy (Gundel et al. 1990, 1993) using coreference-annotated corpora for the languages for which this theory has been originally formulated. Unlike the original, small-scale manual annotations, our approach relies on publicly available corpora with coreference annotations, allowing for statistically significant analyses.

For English and across languages, we confirmed strong associations between **in focus** and the use

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
single-token names							
frequency	5433	6839	6444	n/a	4851	5	23572
correlation	-0,119+++	0,034+++	0,126+++		0,11+++	-0,004 n.s	
multi-token names							
frequency	3582	5006	5531	4389	4765	3	23276
correlation	-0,188+++	-0,042+++	0,083+++	0,147+++	0,108+++	-0,006(+)	
other							
frequency	41882	24654	11895	8758	8058	48	95295
total	50897	36499	23870	13147	17674	56	142143

Table 5: Distribution of names and approximative cognitive statuses in the OntoNotes corpus compared to other referring expressions. Colors are used in accordance with Tab. 1.

of pronouns and zero anaphora, whereas all other types of referring expressions are negatively correlated with **in focus**. Demonstrative pronouns are associated with locally evoked (**activated**) referents; demonstrative NPs showed similar trends, but medial and distal forms lacked consistent associations with cognitive statuses. This suggests that such forms serve specialized discourse functions (e.g., deixis, exophora, contrast) beyond simple givenness marking and should be analyzed separately from the hierarchy. Definite NPs (or Ø NPs in languages lacking definite determiners) tend to occupy the middle-to-lower part of the hierarchy. However, the inability to reliably distinguish **unique** from **referential** and **type** using coreference data makes it difficult to confirm whether definite NPs require **unique** status. While Gundel et al. confirmed this link experimentally, it cannot be directly replicated using anaphoric annotation alone. At the lower end of the hierarchy, indefinite NPs (or Ø NPs in relevant languages) dominate. The absence of previous mention (the primary criterion for **familiar**) helps distinguish them from higher statuses.

Aside from issues with **unique** and medial/distal demonstrative NPs, we confirm GH predictions for Arabic, Chinese, English, Japanese, Korean, Russian, and Spanish. Despite variation in corpus formats, genres, and annotation schemes, and a considerable noise arising from the incomplete nature of coreference annotations in comparison to Gundel et al. (2006), we observed correlations sufficiently strong to yield significant results also for aspects of the Givenness Hierarchy previously described with insufficient amounts of data, only.

Because of difficulties surrounding the approximation of **unique**, we recommend simplifying the GH by collapsing it with **referential** into a novel combined category: **XREF** (extended referential), encompassing both **unique** and **referential** referents as defined by the original GH manual, with the primary criterion for their identification drawn from the GH definition of **referential**. With **XREF**, future studies may better capture the transitional space between **familiar** (hearer-old) and **type identifiable** ($\sim$ hearer-new).

Beyond evaluating the Givenness Hierarchy, we showed how phenomena not originally covered by the Givenness Hierarchy (the givenness of full and short names) can be investigated with this methodology, leading to insights consistent with previous qualitative analyses (Mulkern, 2011). This methodology can thus *infer and extend* language-specific givenness hierarchies. Also, our ability to approximate givenness from coreference annotations is practically relevant: The procedure introduced in Section 2 can bootstrap givenness annotations and thus yield the first available training data for GH annotations. Such data can support the development of automated taggers and serve as evaluation material for future methods of inducing or predicting givenness – potentially even in the absence of coreference annotation. Note that labelling cognitive statuses appears simpler and more robust than full anaphora resolution, while still providing valuable discourse-level insights. Thus, extrapolating GH annotations from coreference annotation may serve as a useful intermediary task for applications requiring discourse-aware processing.

We would like to emphasize that – at this stage – we do not aim to evaluate the theory per se. Although many of the categorizations put forward by Gundel et al. seem plausibile from a cognitive-linguistic perspective, and the factors they consider (proximity, previous mention, assumed hearer knowledge, intention to refer to a specific entity) certainly play a role, it is not uncontested that information is, in fact, categorial by nature (Ariel, 1990), and how these categories are differentiated,

cf. Chiarcos (2010, 2011b,c) for overview, discussion and criticism, and (Poesio and Modjeska, 2008), (Chiarcos, 2011a) or (Hou, 2021) (among others) for possible alternatives and their operationalization. Instead, our primary goal is to demonstrate the potential for using coreference-annotated corpora for bootstrapping Givenness Hierarchies for other languages, and these, in turn, may be useful for the empirical, cross-linguistic evaluation of the theory, potentially, along with other theories of reference, accessibility and information status. We are aware that the approximations suggested in this paper are, to a large extent, imperfect, but we would argue that these heuristics nevertheless capture prototypical examples at sufficient numbers, and that statistically significant patterns observed over these allow us to gain insights into the underlying theoretical models. At this stage, we thus conclude that the theory is to a large extent verifiable with coreference-annotated corpora, that there are limitations in the heuristic identification of UNIQUE referents (for which we suggest a simplification for empirical studies that is based on the implicative nature of the Givenness Hierarchy), but also that not all the claims of the Givenness Hierarchy could be confirmed, especially regarding demonstratives, that motivates more extensive research into other languages, as well, and that this research can be conducted with the bootstrapping methodology suggested here.

As a secondary contribution, we converted two corpora – Japanese NTC 1.5 and Korean KoCoNovel – into CorefUD-compatible formats. The conversion scripts and accompanying materials for these and all other corpora considered here are available under an Apache v.2 license from `https://github.com/acoli-repo/givenness-hierarchy`.

Limitations

This study presents an approximative operationalization of the Givenness Hierarchy using coreference annotations. While most cognitive statuses can be approximated reliably, the status **unique** could not be accurately derived from the available data. For this reason, we propose merging **unique** and **referential** into a single category, **XREF**, in future GH implementations.

Note that to facilitate comparability between our numbers and those of Gundel et al. (1990,1993), we limit our analysis to referring expressions studied by Gundel et al., which restricts coverage and means that totals are not calculated over the full set of referring expressions annotated in a corpus, but only to those that fall into categories also considered by Gundel et al. In particular, proper nouns, quantified nouns, pronominal adverbs and possessive NPs were not considered in Sect. 3. With the extension to proper nouns and the replication of Mulkern (2011), the totals in Sect. 4 were extended to cover anaphors and antecedents annotated as `PROPN`, as well.

Also, we are restricted to referring expressions that we can reliably identify in our data, so, while Gundel et al. (1993) distinguished stressed and unstressed pronouns (available in spoken data), the (primarily) written text we operate with does not provide such cues. Likewise, we did not disambiguate between indefinite *this* and proximal demonstratives, because these are identical in form. In particular, because of limitations in the reliable detection of **unique** referents, we could not naively identify non-**unique** *this*-NPs with 'indefinite *this*' and Gundel et al.'s **referential** category. Similarly, our handling of proper names diverges from Mulkern (2011). Whereas Mulkern seems to focus on person names, exclusively, we evaluated all referring expressions annotated as `PROPN`. Our data thus includes organizations, dates, and locations, as well.

A major limitation lies in the comparability of the corpora. They differ in genre, size, annotation design, and coverage. Crucially, none provide annotations for event anaphora, resulting in underrepresentation of certain referential types (e.g., demonstrative pronouns for events). Annotations for bridging or other models of information status, which could also inform GH annotations, are available for a subset of corpora considered and have been excluded, so that cross-linguistic comparison could be established. Several corpora are also affected by biases in their annotation. For example, OntoNotes and ParCorFull only annotate specific (i.e. **referential** or higher) referents, thus systematically excluding **type identifiable** expressions. Similarly, the KoCoNovel corpus annotates protagonists, only. None of the corpora provide annotations for event anaphora. Russian and Korean corpora lack annotations for zero anaphora (Ø), causing misclassification of referents that should be **in focus** or **activated** as lower-status categories like **familiar**. In Japanese (NTC 1.5), missing text boundaries can lead to erroneous coreference

chains. In the Chinese OntoNotes corpus, an extraction error led to misclassifying a group of proper names as bare nominals.

In order to categorize referring expressions and grammatical features in a systematic way, we rely on Universal Dependencies (UD). While such annotations were available for subset of 5 corpora drawn from the CorefUD collection, they had to be created automatically for Arabic, Chinese, Japanese, the Korean KoCoNovel and the English OntoNotes corpus. We used spaCy (resp., for Arabic, spaCy-udpipe, i.e., UDPipe 2.5) to create them automatically, introducing potential parsing errors as an additional source of noise. For identifying **in focus** referents on the basis of their realization in the preceding sentence (i.e., referents not mentioned in the penultimate sentence, as well), we rely on the UD label `nsubj`, whereas the original definition by Gundel et al. (2006) would also include morphosyntactic topic and focus markers as present (but not annotated) in Korean and Japanese.

Overall, while these limitations introduce variability, our methodology provides a scalable, corpus-based framework for exploring the Givenness Hierarchy across languages and modalities. Despite the aforementioned issues, our analysis is strengthened by recurring patterns observed across multiple languages and multiple independently annotated corpora. We report only statistically significant findings and emphasize robust patterns across datasets, rather than isolated results.

References

Ingrid Abisambra Miccheli and Margaret Lubbers Quesada. 2023. Cognitive status and subject reference in spanish written discourse. *Studies in Hispanic and Lusophone Linguistics*, 16(1):5–33.

M. Ariel. 1990. *Accessing Noun-Phrase Antecedents*. Routledge.

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of pragmatics*, 16(5):443–463.

Keira Gebbie Ballantyne. 2004. Givenness as a ranking criterion in centering theory: evidence from yapese. *Oceanic Linguistics*, 43(1):49–72.

Stefan Baumann and Arndt Riester. 2013. Coreference, lexical givenness and prosody in german. *Lingua*, 136:16–37.

Sarah E Blackwell and Margaret Lubbers Quesada. 2012. Third-person subjects in native speakers' and l2 learners' narratives: Testing (and revising) the givenness hierarchy for spanish. In *Selected proceedings of the 14th Hispanic linguistics symposium*, pages 142–164. Cascadilla Proceedings Project Somerville, MA.

Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

Richard Challoner. 1749-1752 (revision). Douay-Rheims Bible Online. https://www.drbo.org/, accessed 2025-08-05. Official Catholic Version with Search.

Christian Chiarcos. 2010. *Mental Salience and Grammatical Form*. Ph.D. thesis, Potsdam: Universität Potsdam.

Christian Chiarcos. 2011a. Evaluating salience metrics for the context-adequate realization of discourse referents. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 32–43.

Christian Chiarcos. 2011b. The Mental Salience Framework: Context-adequate generation of referring expressions. In Christian Chiarcos, Berry Claus, and Michael Grabski, editors, *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, volume 227, pages 105–140. Walter de Gruyter.

Christian Chiarcos. 2011c. On the dimensions of discourse salience. *Bochumer Linguistische Arbeitsberichte*, 3:31–44.

Logan Daigler, Mark Higger, Terran Mott, and Tom Williams. 2024. Challenges in annotating gesture-based cognitive status in human-robot collaboration datasets. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 364–368.

John Nelson Darby. 1890. Bible in English. Darby Translation. https://kingdomjc.com/Bible_EN.htm, accessed 2025-08-05.

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska, Syahidah Asma Umniyati, and Helena Rodrigues Menezes de Oliveira Vaz. 2024. A multilingual parallel corpus for coreference resolution and information status in the literary domain. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 55–64, Hamburg,Germany. Association for Computational Linguistics.

Talmy Givón. 1983. *Topic Continuity in Discourse: A Quantitative Cross-language Study*, volume 3. John Benjamins Publishing.

Barbara J Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Jeanette Gundel, Nancy Hedberg, Ron Zacharski, Ann Mulkern, Tonya Custis, Bonnie Swierzbin, Amel Khalfoui, Linda Humnick, Bryan Gordon, Mamadou Bassene, and Shana Watters. 2006. Coding protocol for statuses on the givenness hierarchy (gundel, hedberg and zacharski 1993). Technical report, University of Minnesota.

Jeanette K Gundel. 1998. Centering theory and the givenness hierarchy: Towards a synthesis. *Centering theory in discourse*, pages 183–198.

Jeanette K Gundel, Mamadou Bassene, Bryan Gordon, Linda Humnick, and Amel Khalfaoui. 2010. Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, 42(7):1770–1785.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1990. Givenness, implicature, and the form of referring expressions. In *Annual Meeting of the Berkeley Linguistics Society*, pages 442–453.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307. Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society (1990), pp. 442-453.

Jeanette K Gundel and Kaitlin Johnson. 2013. Children's use of referring expressions in spontaneous discourse: Implications for theory of mind development. *Journal of Pragmatics*, 56:43–57.

Zhao Han, Polina Rygina, and Thomas Williams. 2022. Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th international conference on natural language generation*, pages 1–14.

Nancy Hedberg. 2010. Centering and noun phrase realization in kaqchikel mayan. *Journal of Pragmatics*, 42(7):1829–1841.

Nancy Hedberg and Nathalie Schapansky. 1996. A referentiality constraint on preverbal nps in breton. In *annual meeting of the LSA, San Diego, Jan*, pages 3–7.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Mark Higger and Tom Williams. 2024. Gaia: A givenness hierarchy theoretic model of situated referring expression generation. In *Proceedings of the annual meeting of the cognitive science society*, volume 46.

Yufang Hou. 2021. End-to-end neural information status classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1377–1388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linda Humnick. 2005. Pronominal references and the encoding of cognitive status in kumyk. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 41, pages 149–163. Chicago Linguistic Society.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. 2017. Naist text corpus: Annotating predicate-argument and coreference relations in japanese. In *Handbook of linguistic annotation*, pages 1177–1196. Springer.

Toldova S Ju, A Roytberg, AA Ladygina, MD Vasilyeva, IL Azerkovich, M Kurzukov, G Sim, DV Gorshkov, A Ivanova, A Nedoluzhko, and 1 others. 2014. Rueval-2014: Evaluating anaphora and coreference resolution for russian. *Computational Linguistics and Intellectual Technologies*, 13:681–694.

Alireza Khormaee and Elina Skrouchi. 2015. A study of discourse anaphora in persian language based on gundel, hedberg and zacharski's givenness hierarchy. *Language Science*, 3(4):154–111.

Hyunwoo Kim. 2016. The effects of accessibility of english reference in korean efl learners' sentence processing. *ITL-International Journal of Applied Linguistics*, 167(1):78–97.

Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. Koconovel: Annotated dataset of character coreference in korean novels. *Preprint*, arXiv:2404.01140.

Martina Krüger. 2018. *Prosodic decoding and encoding of referential givenness in adults with autism spectrum disorders*. Ph.D. thesis, Universität zu Köln.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. Parcorfull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ann E Mulkern. 2008. Knowing who's important: Relative discourse salience and irish pronominal forms. In *The Grammar–Pragmatics Interface: Essays in honor of Jeanette K. Gundel*, pages 113–142. John Benjamins Publishing Company.

Ann E Mulkern. 2011. The game of the name. In *Reference and referent accessibility*, pages 235–250. John Benjamins Publishing Company.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Achuodho V Omondi. 2016. *Topic and Focus in Dholuo Oral Speeches: Case of Givenness Hierarchy Framework.* Ph.D. thesis, University of Nairobi.

Poulomi Pal, Grace Clark, and Tom Williams. 2021. Givenness hierarchy theoretic referential choice in situated contexts. In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

Massimo Poesio and Natalia N Modjeska. 2008. Focus, activation, and this-noun phrases: An empirical study. In *Anaphora processing: Linguistic, cognitive and computational modelling*, pages 429–449. John Benjamins Publishing Company.

Sameer S Pradhan and Nianwen Xue. 2009. Ontonotes: the 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*.

Marta Recasens and M Antònia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language resources and evaluation*, 44(4):315–345.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: an evaluation across different types of texts. In *LREC*.

Petr Sgall, Eva Hajicová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Kevin Spevak, Zhao Han, Tom Williams, and Neil T Dantam. 2022. Givenness hierarchy informed optimal document planning for situated human-robot interaction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6109–6115. IEEE.

Laurie A Swabey. 2011. Referring expressions in asl discourse. *Discourse in signed languages*, pages 96–120.

Elizabeth Closs Traugott and Susan Pintzuk. 2008. Coding the york-toronto-helsinki parsed corpus of old english prose to investigate the syntax-pragmatics interface. *Studies in the History of the English Language IV. Empirical and Analytical Advances in the Study of English Language Change. Berlin/New York: Mouton de Gruyter*, pages 61–80.

Marta Velnic. 2018. *Ditransitive structures in Croatian adult and child language: The role of animacy and givenness*. Ph.D. thesis, UiT Norges arktiske universitet.

Weijie Xu and Ming Xiang. 2021. Is there a predictability hierarchy in reference resolution? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Appendix

In this appendix, we provide the detailed list of referring expressions (Tab. 6) considered by us and Gundel et al. (1990, 1993), as well as raw frequencies and correlation analyses for the individual corpora. Table 3 aggregates over Tables 8, 9, 10 and 11 for English. Table 4 aggregates over Tab. 12 for Arabic, Tab. 13 for Chinese, Tab. 14 for Japanese, Tables 15 and 16 for Korean, Tab. 17 for Russian and Tab. 18 for Spanish. As for the color codes, we use the same schema as Tab. 1.

	Arabic	Chinese	English	Japanese	Korean	Russian	Spanish
Ø	Ø	Ø	[0]	Ø	(Ø)[1]	(Ø)	Ø
pron	*هُوَ ، هِيَ ، هُم*	*他, 他们, 她*	*he, she, it*	*彼, 彼女*	[2]	*он, она, они*	*él*
Gundel&al label	hua[3]	tā	it	kare		on	él
dem.prox	*هٰذِي ، هٰذَيْنِ*	*这, 这个*	*this, these*	*これ*	*이*	*это*	*éste*
Gundel&al label	haaðaa[3]	zhè	this	kore	i[3]	èto	éste
dem.med				*それ*	*그*		*todo ese*
Gundel&al label				sore	kū[3]		ése
dem.dist	*أُولَئِكَ ، تِلْكَ ، تِيكَ*	*那, 那个*	*that, those*	*あれ*	*저*	*то*	*todo aquel*
Gundel&al label	ðaalika[3]	nèi	that	are	cə[3]	to	aquél
dem.prox N	*-هٰؤُلاءِ المُعارِضِينَ*	*这次会议*	*these people*	*この事件*	*이런 식으로*	*этот процесс*	*este situación*
Gundel&al label	haaðaa N[3]	zhè N	this N	kono N	i N[3]	èto N	este N
dem.med N				*その責任*	*그 곳에서*		*ese día*
Gundel&al label				sono N	kū N[3]		ese N
dem.dist N	*أُولَئِكَ الحِيتانِ وَ- الدينا صورات-*	*那句话*	*that day*	*あの戦争*	*저 양반*	*тот явление*	*aquel país*
Gundel&al label	ðaalika N[3]	nèi N	that N	ano N	cə N[3]	to N	aquel N
def N	*اللَجْنَةِ*		*the company*				*el Gobierno*
Gundel&al label	al N[3]		the N				el N
Ø N[4]	Ø N	Ø N		Ø N	Ø N	Ø N	Ø N
a (one) N		*一个人*	*a year*				*uno año*
Gundel&al label		yi N	a N				un N

[0] Non-applicable categories rendered in gray
[1] Referring expressions in brackets are not covered by the annotation.
[2] We follow Gundel et al. (1990) in considering Korean 그 a proximal demonstrative, not a personal pronoun.
[3] According to Gundel et al. (1990). Unmarked labels follow Gundel et al. (1993).
[4] Only for languages for which *Ø N* was also considered by Gundel et al. (1990, 1993).

Table 6: Inventories of referring expressions considered by Gundel et al. (1990, 1993) and in this study, listing type of referring expression, original label used by Gundel et al. and corpus examples

	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

Table 7: Cognitive statuses and referential expressions in English as reported by Gundel et al. (1993), absolute numbers, for correlation and binary χ^2 significance see Tab. 1

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	33056	10695	1983		812	24	46570
dem.prox	325	881	199		204	1	1610
dem.dist	606	1236	157		144	6	2149
dem.prox N	688	1421	871	226	532	5	3743
dem.dist N	450	750	238	315	151		1904
the N	6217	9435	8211	5630	4888	11	34392
a N	540	236	236	2587	1327	1	4927
total	41882	24654	11895	8758	8058	48	95295

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,532+++	-0,065+++	-0,243+++	-0,311+++	-0,236+++	0,001 n.s
dem.prox	-0,063+++	0,086+++	0 n.s	-0,042+++	0,02+++	0,001 n.s
dem.dist	-0,048+++	0,11+++	-0,024+++	-0,048+++	-0,01+	0,015+++
dem.prox N	-0,104+++	0,056+++	0,066+++	-0,022+++	0,042+++	0,007(+)
dem.dist N	-0,058+++	0,044+++	0 n.s	0,036+++	-0,003 n.s	-0,003 n.s
the N	-0,392+++	0,027+++	0,259+++	0,187+++	0,155+++	-0,006 n.s
a N	-0,155+++	-0,112+++	-0,054+++	0,35+++	0,155+++	-0,003 n.s

Table 8: Referring expressions in OntoNotes 5.0, English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	3066	949	258	7	323	117	4720
dem.prox	105	162	38		72	39	416
dem.dist	117	174	42	1	35	34	403
dem.prox N	155	216	156	130	109	216	982
dem.dist N	101	76	19	54	34	90	374
the N	1046	910	1082	1169	1021	2921	8149
a N	691	130	123	731	416	1628	3719
total	5281	2617	1718	2092	2010	5045	18763

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,475+++	0,103+++	-0,074+++	-0,203+++	-0,073+++	-0,319+++
dem.prox	-0,01 n.s	0,109+++	0 n.s	-0,053+	0,032+++	-0,059+++
dem.dist	0,003 n.s	0,125+++	0,007 n.s	-0,051+++	-0,01 n.s	-0,062+++
dem.prox N	-0,065+++	0,055+++	0,055+++	0,016(+)	0,003 n.s	-0,026++
dem.dist N	-0,004 n.s	0,026++	-0,02+	0,015(+)	-0,007 n.s	-0,009 n.s
the N	-0,298+++	-0,07+++	0,125+++	0,089+++	0,051+++	0,177+++
a N	-0,106+++	-0,15+++	-0,101+++	0,134+++	0,008 n.s	0,189+++

Table 9: Referring expressions in the GUM corpus (UDcoref 1.3), English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	6488	1233	298		132	90	8241
dem.prox		1	2		1	1	5
dem.dist					1	1	2
dem.prox N	35	53	43	6	25	41	203
dem.dist N	43	22	16	14	17	36	148
the N	375	409	710	82	456	1329	3361
a N	110	43	41	80	225	883	1382
total	7051	1761	1110	182	857	2381	13342

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,659+++	0,066+++	-0,216+++	-0,149+++	-0,25+++	-0,556+++
dem.prox	-0,02n/a	0,004n/a	0,022n/a	-0,002n/a	0,011n/a	0,001 n/a
dem.dist	-0,013 n/a	-0,005n/a	-0,004n/a	-0,001n/a	0,022n/a	0,01n/a
dem.prox N	-0,089+++	0,047+++	0,058+++	0,017(+)	0,03++	0,008 n.s
dem.dist N	-0,05+++	0,005 n.s	0,01 n.s	0,074 n/a	0,022(+)	0,018(+)
the N	-0,485+++	-0,018(+)	0,269+++	0,054+++	0,169+++	0,329+++
a N	-0,306+++	-0,101+++	-0,066+++	0,13+++	0,137+++	0,409+++

Table 10: Referring expressions in the LitBank corpus (UDcoref 1.3), English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	195	66	3		1	11
dem.prox	6	9	2			1
dem.dist	12	4				
dem.prox N	3	8	1		2	
dem.dist N		1		1	2	
the N	14	17	9	1	29	1
a N		1	1	1	12	
total	230	106	16	3	46	13

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,43+++	-0,055 n.s	-0,204+++	-0,121 n/a	-0,484+++	0,069 n/a
dem.prox	-0,095 n.s	0,119(+)	0,08 n/a	-0,018 n/a	-0,075 n/a	0,03 n/a
dem.dist	0,078 n.s	-0,003 n/a	-0,04 n/a	-0,017 n/a	-0,071 n/a	-0,036 n/a
dem.prox N	-0,128+	0,135 n/a	0,032 n/a	-0,016 n/a	0,019 n/a	-0,034 n/a
dem.dist N	-0,11 n/a	-0,001 n/a	-0,02 n/a	0,283 n/a	0,122 n/a	-0,018 n/a
the N	-0,328+++	-0,017 n.s	0,208+++	0,037 n/a	0,43+++	-0,045 n/a
a N	-0,217++	-0,084 n/a	0,028 n/a	0,136 n/a	0,425 n/a	-0,035 n/a

Table 11: Referring expressions in the ParCorFull corpus (UDcoref 1.3), English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	3651	684	99	3	53	1	4491
pron	521	34	4		7		566
dem.prox	1	1			1		3
dem.dist					1		1
dem.prox N			1		1		2
dem.dist N				3			3
the N	1780	2168	1379	1391	2714	1	9433
N	401	785	636	1780	1761		5363
total	6354	3672	2119	3177	4538	2	14499

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,506+++	-0,156+++	-0,235+++	-0,354+++	-0,435+++	0,005 n/a
pron	0,196+++	-0,09+++	-0,079+++	-0,107+++	-0,131+++	-0,002 n/a
dem.prox	-0,003 n/a	0,003 n/a	-0,006 n/a	-0,008 n/a	0,001 n/a	0 n/a
dem.dist	-0,007 n/a	-0,005 n/a	-0,003 n/a	-0,004 n/a	0,012 n/a	0 n/a
dem.prox N	-0,01 n/a	-0,007 n/a	0,012 n/a	-0,006 n/a	0,005 n/a	0 n/a
dem.dist N	-0,013 n/a	-0,008 n/a	-0,006 n/a	0,027 n/a	-0,01 n/a	0 n/a
the N	-0,686+++	-0,074+++	0 n.s	-0,236+++	-0,074+++	-0,004 n/a
N	-0,561+++	-0,188+++	-0,06+++	0,209+++	0,025+	-0,009 n/a

Table 12: Referring expressions in OntoNotes 5.0, Arabic

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	11395	3403	1221		1121		17140
pron	7509	2416	688	179	437	1	11230
dem.prox	659	576	278				1513
dem.dist	184	147	87	83	112		613
dem.prox N	719	703	351	78	209		2060
dem.dist N	78	78	63	37	52		308
N	34344	17768	14701	12512	20002	7	99334
a N	209	210	173	900	630		2122
total	55097	25301	17562	13789	22563	8	132198

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,194+++	0,007+	-0,07+++	-0,132+++	-0,108+++	-0,003 n/a
pron	0,156+++	0,018+++	-0,064+++	-0,088+++	-0,107+++	0,001 n/a
dem.prox	0,004 n.s	0,052+++	0,016+++	-0,037+++	-0,049+++	-0,001 n/a
dem.dist	-0,016+++	0,008+	0,002 n.s	0,007(+)	0,002 n.s	-0,001 n/a
dem.prox N	-0,017+++	0,048+++	0,014+++	-0,027+++	-0,023+++	-0,001 n/a
dem.dist N	-0,016+++	0,008+	0,01++	0,003 n.s	0 n.s	0 n/a
N	-0,25+++	-0,055+++	0,078+++	0,123+++	0,142+++	0,002 n/a
a N	-0,082+++	-0,03+++	-0,019+++	0,134+++	0,043+++	-0,001 n/a

Table 13: Referring expressions in OntoNotes 5.0, Chinese

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	25516	7240	4790		99	1	37646
pron	42	36	15			1	94
dem.prox	72	213	414		10	2	711
dem.med	121	178	406		20	5	730
dem.dist		6	1				7
dem.prox N	150	367	557	1	10	3	1088
dem.med N	69	108	578		18	5	778
dem.dist N	2	5	19		1		27
N	12600	13951	100305	138	3034	927	130955
total	38572	22104	107085	139	3192	944	172036

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,576+++	0,101+++	-0,541+++	-0,015+	-0,062+++	-0,039+++
pron	0,012+++	0,018+++	-0,022+++	-0,001 n/a	-0,003 n/a	0,002 n/a
dem.prox	-0,019+++	0,033+++	-0,005(+)	-0,002 n/a	-0,002 n.s	-0,002 n/a
dem.med	-0,009++	0,023+++	-0,009++	-0,002 n/a	0,004 n.s	0,001 n/a
dem.dist	-0,003 n/a	0,014 n/a	-0,006 n/a	0 n/a	-0,001 n/a	0 n/a
dem.prox N	-0,017+++	0,05+++	-0,018+++	0 n/a	-0,006(+)	-0,003 n.s
dem.med N	-0,022+++	0,002 n.s	0,017+++	-0,002 n/a	0,002 n.s	0,001 n/a
dem.dist N	-0,005 n.s	0,002 n/a	0,002 n.s	0 n/a	0,002 n/a	-0,001 n/a
N	-0,548+++	-0,117+++	0,528+++	0,015+++	0,061+++	0,038+++

Table 14: Referring expressions in NTC 1.5, Japanese

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
dem.prox	106	58	21			565	750
dem.med	1391	897	365			312	2965
dem.dist			1			1	2
dem.prox N	68	87	23	18		336	532
dem.med N	34	21	8	30	1	376	470
dem.dist N						1	1
N	3465	3396	3385	178	4782	25642	40848
subtotal	5064	4459	3803	226	4783	27233	45568

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
dem.prox	0,012+	-0,009 n.s	-0,026+++	-0,009 n/a	-0,044+	0,041+++
dem.med	0,301+++	0,182+++	0,038+++	-0,019 n.s	-0,09+++	-0,265+++
dem.dist	-0,002 n/a	-0,002 n/a	0,01 n/a	0 n/a	-0,002 n/a	-0,001 n/a
dem.prox N	0,006 n.s	0,024+++	-0,016++	0,045 n/a	-0,037+	0,008 n.s
dem.med N	-0,013+	-0,018+++	-0,025+++	0,086 n/a	-0,034+++	0,042+++
dem.dist N	-0,002 n/a	-0,002 n/a	-0,001 n/a	0 n/a	-0,002 n/a	0,004 n/a
N	-0,246+++	-0,146+++	-0,006 n.s	-0,025+++	0,116+++	0,181+++

Table 15: Referring expressions in the ECMT corpus (UDcoref 1.3), Korean

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
dem.prox		4	10		4	6	24
dem.med	631	710	255		167	77	1840
dem.dist	42	21	12		10	1	86
dem.prox N	34	19	12		8	7	80
dem.med N	43	32	22		20	15	132
dem.dist N	5	6	2		2	2	17
N	1812	2391	1670	1	1357	825	8056
total	2567	3183	1983	1	1568	933	10235

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
dem.prox	-0,028 n.s	-0,015 n.s	0,027+	0 n/a	0,002 n.s	0,027+
dem.med	0,103+++	0,076+++	-0,066+++	-0,005 n/a	-0,082+++	-0,081+++
dem.dist	0,05+++	-0,013 n.s	-0,013 n.s	-0,001n/a	-0,009 n.s	-0,025(+)
dem.prox N	0,036++	-0,014 n.s	-0,01 n.s	-0,001 n/a	-0,013 n.s	-0,001 n.s
dem.med N	0,02(+)	-0,017 n.s	-0,008 n.s	-0,001 n/a	-0,001 n.s	0,009 n.s
dem.dist N	0,004 n/a	0,004 n.s	-0,008 n/a	0 n/a	-0,004 n/a	0,004 n/a
N	-0,108+++	-0,062+++	0,065+++	0,005 n/a	0,081+++	0,072+++

Table 16: Referring expressions in the KoCoNovel corpus, Korean

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	1253	543	71		30	13	1910
dem.prox	2	1			1		4
dem.dist							
dem.prox N	86	129	39		42	7	303
dem.dist N		6	5		30		41
N	848	950	1062	26	1825	60	4771
total	2189	1629	1177	26	1928	80	7029

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,455+++	0,076+++	-0,213+++	-0,037 n.s	-0,354+++	-0,026(+)
dem.prox	0,01 n/a	0,001 n/a	-0,011 n/a	-0,001 n/a	-0,001 n/a	-0,003 n/a
dem.prox N	-0,013 n.s	0,098+++	-0,022 n.s	-0,013 n.s	-0,065+++	0,023 n/a
dem.dist N	-0,052(+)	-0,016 n.s	-0,009 n.s	-0,005 n/a	0,079+++	-0,008 n/a
N	-0,42+++	-0,112+++	0,215+++	0,042++	0,353+++	0,016 n.s

Table 17: Referring expressions in RuCor (UDcoref 1.3), Russian

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	6098	851	197	8	215	24	7393
pron	1536	313	42	25	129	14	2059
dem.prox	100	79	10	2	20	3	214
dem.med	65	72	4	1	5		147
dem.dist	4	3	1		2	1	11
dem.prox N	471	444	153	88	105	14	1275
dem.med N	214	181	32	39	22	3	491
dem.dist N	23	7	2	5	4		41
the N	5503	4256	3098	1285	4691	159	18992
a N	270	85	64	37	328	8	792
total	14014	6206	3539	1453	5193	218	30623

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,416+++	-0,123+++	-0,157+++	-0,123+++	-0,211+++	-0,026+++
pron	0,155+++	-0,034+++	-0,08+++	-0,045+++	-0,077+++	-0,001 n.s
dem.prox	0,002 n/a	0,035+++	-0,018++	-0,015++	-0,017++	0,007 n/a
dem.med	-0,002 n.s	0,05+++	-0,019++	-0,013(+)	-0,025+++	-0,006 n/a
dem.dist	0,019 n.s	0,003 n/a	-0,001 n/a	-0,004 n/a	0,001 n/a	0,019++
dem.prox N	-0,037+++	0,075+++	0,003 n.s	0,021++	-0,048+++	0,01 n.s
dem.med N	-0,006 n.s	0,053+++	-0,02++	0,019++	-0,042+++	-0,002 n/a
dem.dist N	0,008 n.s	-0,003 n.s	-0,008 n.s	0,013(+)	-0,007 n.s	-0,003 n.s
the N	-0,431+++	0,068+++	0,19+++	0,121+++	0,264+++	0,019++
a N	-0,038+++	-0,039+++	-0,018+	-0,001 n.s	0,106+++	0,006 n.s

Table 18: Referring expressions in AnCora (UDcoref 1.3), Spanish