The Narrative Niche: Communicative Need in Narrative Drives Lexical Richness within Efficient Communication

Jules Torgue[1,2], **Francis Mollica**[2] **& Barend Beekhuizen**[1]
[1]Department of Linguistics, University of Toronto, Canada
[2]Melbourne School of Psychological Sciences, University of Melbourne, Australia

Abstract

Efficient communication accounts of cross-linguistic variation propose that languages trade-off informativeness and complexity based on communicative need, often measured by usage frequency and operationalized as a free parameter. We propose that this need is not uniform across discursive genres and that the narrative genre applies a stronger pressure for lexical informativeness than non-narrative genres. Using spoken-language data from typologically diverse languages, we analyze the informativeness of lexical fields in narrative and other genres, as operationalized by their lexical richness. We show that frequency in narrative genres predicts lexical richness more strongly than in other genres. This supports the theory that narrative is a privileged communicative context that drives the informativity displayed by different parts of the lexicon.

Keywords: communicative efficiency, semantic typology, corpus linguistics, narrative

Introduction

Languages across the world vary in the way they partition meaning through their lexicon. Where English makes a distinction between the experience of *seeing*, the activity of *looking*, and the extended activity of *watching*, French compresses the same meanings into *voir*, translating *see*, and *regarder*, translating *look* and *watch*. Here, we can say that French **colexifies** (François, 2008; Haspelmath, 2023) the 'look' and 'watch' meanings that English **dislexifies**. Such variation across languages has been successfully explained through the idea that languages are shaped by a general principle of efficient communication (Kemp et al., 2018). Formal accounts based on this idea propose that pressures for efficient communication drive languages towards an optimal trade-off between the complexity (more complex systems are harder to remember and transmit) and informativeness (more complex systems can allow for more precise communication) of semantic systems (Kemp et al., 2018; Zaslavsky et al., 2018).

The position on this optimal frontier is governed by the relative weight a language assigns to informativity, compared to complexity. This weight is thought to be governed by the communicative need of the semantic domain, with greater communicative need permitting systems to be more complex so that they can be more informative (Zaslavsky et al., 2020; Bradford et al., 2022). The notion of communicative need has been made concrete in various ways, as we will explore below. One of these ways is the frequency of usage of a semantic domain: if we talk more about a semantic domain, we are more likely to "accept" the cost of greater complexity in return for higher informativity – i.e., a lower (absolute) number of misunderstandings. Crucially, the realization that there is some parameter that modulates the informativity-complexity trade-off opens the door to further empirical exploration of what goes into it. Here, we inquire into the **usage frequency** of a semantic domain as a factor modulating the informativity-complexity trade-off. In particular, we consider how the usage frequency of domains may impact the degree of lexical informativity differently for different linguistic activities.

Language users engage in different types of **linguistic activities** (Levinson, 1979), longer stretches of discourse that have goals and constraints on the nature of the language that is found in them, and whose labels are the answer to the question "what are we doing?". Many activity types only exist in certain cultures (e.g., the service encounter or the romance novel) whereas others are arguably universal: for instance, narratives, dyadic goal-oriented conversations, instructions and other forms of joint deliberation (Bakhtin, 1986). Here we ask whether usage frequency modulates trade-off between informativity and complexity similarly across activities.

We propose that the linguistic activity of narrative has a special status, and that usage frequency in narrative can be expected to be more impactful than usage frequency in other (universal) activities. Narrative has been identified as a communicative and cognitive context that plays a special role in the (cultural) evolution of language and that has genre-related properties favouring informativity, as discussed below. For both narrative and non-narrative genres, greater usage frequency is expected to lead to greater lexical richness, but we expect this effect to be stronger for narrative. After motivating this hypothesis, we demonstrate its validity on the basis of a typologically diverse corpus consisting of spoken-language data in 26 languages. This works forms a first step of a larger line of work: demonstrating empirically that there is a non-uniformity in the impact of usage frequency across linguistic activities, and more generally, grounding the design properties of lexicons more extensively in properties of the contexts of use. In the Discussion, we will pick up questions about potential formalization and further theoretical integration.

Background

In the literature on efficient communication, the communicative need of semantic domains and concepts in such domains has been derived in a number of different ways. In the Infor-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

mation Bottleneck approach, the β parameter is simply fit to produce the closest-to-optimal result (Zaslavsky et al., 2018), while domain-internal variation in need ("object-level need") may be estimated from linguistic usage, properties of the environment (Gibson et al., 2017), or a combination (Zaslavsky et al., 2020). Other work has explored the estimation of domain (and object) need probabilities from text corpora: Bradford et al. (2022) statistically model the degree of lexical precision of languages in the CLICS^3 dataset (Rzymski et al., 2020) on the basis of language-invariant domain-need estimates from text corpora in seven languages, but do not address how variation *between* languages may play a role. This latter point is, in turn, taken up by Anand & Regier (2023), who show that the usage frequencies of specific kinship concepts (i.e., object-level need) track with the informativeness in specific parts of the domain. Similarly, Beekhuizen (2024) demonstrates that the variation in domain-level need between languages correlates positively with the lexical richness.

Here, we explore the idea that some *kinds of usage* may affect the trade-off between complexity and informativeness more than others. Several facets of the usage situation may play into this variable 'weighing' of usage events, but a salient first candidate to explore is the **linguistic activity** (Levinson, 1979). Levinson characterizes these activities as longer spates of discourse that are oriented towards a goal, constrain the language recruited for that goal, and form the basis of the inferences we draw on the basis of that language. Especially the goal-directness of linguistic activities (and 'genres' – the term often used for the written activities: Swales, 1990) is relevant here.

We argue that some activities have higher semantic stakes than others, that is to say: some activities require from their interlocutors a greater granularity of the **construal** of the situation Langacker (2019), that is: the way in which (real or imagined) states of affairs are presented. This activity-bound demand on granularity may emerge from (referential and inferential) goals of the activity.

We further argue that narrativity is such a 'high stakes' situation, for several reasons pertaining to its activity-bound goals. First, Bruner (1990) delineates three central properties of narratives: (1) they recount a sequence of past events, (2) they involve characters, and (3) these two aspects are organized with respect to the overarching sequence: the plot. The focus on past tense events that are not in the here and now make narrative instantiate the linguistic 'design feature' of **displacement** (Hockett, 1960). Displaced events are not otherwise experientially accessible, and as such, language is the only means for recipients to conceptualize the situation. Second, the presented sequence of events must be **reportable** or **tellable** (Labov & Waletzky, 1967; Labov, 2013) for a narrative to be a good instance of its activity type: they typically start with an ordinary situation that leads to a complication that makes the situation extra-ordinary, and subsequently to a resolution that restores normalcy. Fine-grained conceptual distinctions allow a narrator to accomplish such tellability. Third, narratives typically involve the presentation of the subjective perspectives of the narrator and staged participants ('vantage points'; Labov & Waletzky, 1967) and, as a conclusion, an evaluation of the events that answers questions such as "Was the protagonist morally in the right?" and "What lessons can we learn?" (Labov, 2013). The latter properties underscore the high stakes nature of story telling: through evaluation and subjective perspectives, narratives play a key role in how we position ourselves socially (Schiffrin, 1996) and how we transmit (normative) cultural knowledge (Bietti et al., 2019). Again, fine-grained conceptual distinctions afford the presentation of a version (cf. Cuff, 1993) of the events at hand that leads to the intended moral inferences.

As an activity type, narrative has been argued to be a "primary speech genre" (Bakhtin, 1986), that is: not derived from other speech genres (as many other genres, such as police interrogations and romance novels, are). Bruner (1990) similarly proposes humans are imbued with "a readiness or predisposition to organize experience into a narrative form" (p. 43). This primacy goes hand in hand with its **presumed universality** – to the best of our knowledge, and despite variation in its exact form, no culture lacks the activity type of narrative (though Everett (2010) argues the Pirahã have a strong cultural bias against displacement). Various strands of research have explored potential consequences of this (presumed) universal status. For Tomasello (2010), the emergence of narrative is argued to form a cultural evolutionary niche in which certain aspects of natural language grammar could (culturally) evolve: in particular, definiteness, tense, aspect, and modality. Corballis (2009) identifies 'mental time travel' as the cognitive ability enabling temporal and narrative thinking, and suggests it to be a cognitive substrate which must have driven the emergence of richer forms of language. In the same line as Bruner's proposition, Boyd (2018) and Ferretti & Adornetti (2020) suggest that language and the narrative activity must have co-evolved.

Here, we ask whether narrativity, more so than other linguistic activities, constitutes a cultural evolutionary niche within which the lexicons of languages as a whole develop. Concretely, we predict that, while **greater usage frequency** is expected to **lead to greater informativity** in the form of a richer lexical inventory given a meaning to be expressed, this effect is **stronger for narrative** than for other everyday speech genres, due to the three design features outlined above: displacement, tellability, and moral/social implicativity.

Data and Methods

Data

To assess this hypothesis, we use the DoReCo corpora (Seifart et al., 2024), a collection of spoken language corpora gathered by documentary linguists. We select 48 of the 53 languages that have both data marked as narrative and as non-narrative (instruction, free conversation). Notably, the latter genres are also universal linguistic activities, though ones with, as we argue, lower semantic stakes. Table 1 presents the corpus,

language (glottocode, family, area: reference)	*N* tokens	language (glottocode, family, area: reference)	*N* tokens
Anal (anal1239, Sino-Tibetan, Eurasia: Ozerov, 2024)	20041	Northern Kurdish (Kurmanji) (nort2641, Indo-European, Eurasia: Haig et al., 2024)	18824
Yali (Apahapsili) (apah1238, Nuclear Trans New Guinea, Papunesia: Riesberg, n.d.)	18726	Northern Alta (nort2875, Austronesian, Papunesia: Garcia-Laguia, 2024)	15818
Arapaho (arap1274, Algic, North America: Cowell, 2024)	35839	Fanbyak (orko1234, Austronesian, Papunesia: Franjieh, 2024)	19309
Baïnounk Gubëeher (bain1259, Atlantic-Congo, Africa: Cobbinah, 2024)	22733	Pnar (pnar1238, Austroasiatic, Eurasia: Ring, 2024)	20796
Beja (beja1238, Afro-Asiatic, Africa: Vanhove, 2024)	38121	Resígaro (resi1247, Arawakan, South America: Seifart, 2024b)	11552
Bora (bora1263, Boran, South America: Seifart, 2024a	53905	Daakie (port1286, Austronesian, Papunesia: Krifka, 2024)	14281
Cabécar (cabe1245, Chibchan, North America: Quesada et al., 2024)	8472	Ruuli (ruul1235, Atlantic-Congo, Africa: Witzlack-Makarevich et al., 2024)	20133
Cashinahua (cash1254, Pano-Tacanan, South America: Reiter, 2024)	21875	Sadu (sadu1234, Sino-Tibetan, Eurasia: Xu & Bai, 2024)	12286
Dolgan (dolg1241, Turkic, Eurasia: Däbritz et al., 2024)	34863	Sanzhi Dargwa (sanz1248, Nakh-Daghestanian, Eurasia: Forker & Schiborr, 2024)	19693
Evenki (even1259, Tungusic, Eurasia: Kazakevich & Klyachko, 2024)	19877	Savosavo (savo1255, Isolate, Papunesia: Wegener, 2024)	10897
Goemai (goem1240, Afro-Asiatic, Africa: Hellwig, 2024)	46673	Nafsan (South Efate) (sout2856, Austronesian, Papunesia: Thieberger, 2024)	26969
Gorwaa (goro1270, Afro-Asiatic, Africa: Harvey, 2024)	31186	English (Southern England) (sout3282, Indo-European, Eurasia: Schiborr, 2024)	44201
Hoocąk (hoch1243, Siouan, North America: Hartmann, 2024)	10729	Sümi (sumi1235, Sino-Tibetan, Eurasia: Teo, 2024)	10345
Jahai (jeha1242, Austroasiatic, Eurasia: Burenhult, 2024)	13060	Svan (svan1243, Kartvelian, Eurasia: Gippert, 2024)	19313
Jejuan (jeju1234, Koreanic, Eurasia: Kim, 2024)	14169	Tabasaran (taba1259, Nakh-Daghestanian, Eurasia: Bogomolova et al., 2024)	9346
Kakabe (kaka1265, Mande, Africa: Vydrina, 2024)	53428	Teop (teop1238, Austronesian, Papunesia: Mosel, 2024)	16777
Kamas (kama1351, Uralic, Eurasia: Gusev et al., 2024)	77445	Texistepec Popoluca (texi1237, Mixe-Zoque, North America: Wichmann, 2024)	19218
Tabaq (Karko) (kark1256, Nubian, Africa: Hellwig et al., 2024)	17835	Mojeño Trinitario (trin1278, Arawakan, South America: Rose, 2024)	31221
Komnzo (komn1238, Yam, Papunesia: Döhler, 2024)	47393	Asimjeeg Datooga (tsim1256, Nilotic, Africa: Griscom, 2024)	14529
Lower Sorbian (lowe1385, Indo-European, Eurasia: Bartels & Szczepański, 2024)	13814	Urum (urum1249, Turkic, Eurasia: Skopeteas et al., 2024)	29006
Movima (movi1243, Isolate, South America: Haude, 2024)	17308	Vera'a (vera1241, Austronesian, Papunesia: Schnell, 2024)	23973
Dalabon (ngal1292, Gunwinyguan, Australia: Ponsonnet, 2024)	5183	Yongning Na (yong1270, Sino-Tibetan, Eurasia: Michaud, 2024)	26326
N‖ng (nngg1234, Tuu, Africa: Güldemann et al., 2024)	40876	Yucatec Maya (yuca1254, Mayan, North America: Skopeteas, 2024)	9936
Nisvai (nisv1234, Austronesian, Papunesia: Aznar, 2024)	18217	Yurakaré (yura1255, Isolate, South America: Gipper & Ballivián Torrico, 2024)	8128

Table 1: The 48 selected languages in the DoReCo dataset. '*N* tokens' is the number of target language word tokens.

including citations to used datasets as part of terms of use.

A pipeline for field extraction

To compare the lexical richness of language given the same meaning in a lexicon-wide manner, we need to have a meaning-related unit of comparison that is applicable across all languages (Haspelmath, 2018). Here, we propose to use the free translations into English as the starting point for representing meanings.

Seed Items: In particular, we use the lemmas of open-class words in English as seed items ($N = 6,130$ unique lemmas across the 48 languages, with errors semi-automatically removed) representing individual meanings. We *could* then simply computationally extract the translation equivalents of these seed items into the target languages, and subsequently compare the lexical richness per seed item (e.g., the number of unique terms, their entropy, as explained below). However, this would only allow us to find cases where target languages dislexify ('express as two or more lexical items'; Haspelmath, 2023) what English colexifies. This procedure would miss out on the cases where English dislexifies what target languages colexify. To avoid this English-centered bias, we design a two-step procedure, involving a Colexification step, and a Field Detection step to find fields, sets of meanings (i.e., English lemmas), that are regularly colexified across languages.

Retrieving Translation Equivalents (TEs): The input to the two-step procedure consists of a mapping, for each language, of each English seed item (in the free translations) to automatically stemmed word forms (**stems**) in the target language alongside the specific **word forms** belonging to that stem. This component is necessary to ensure comparability of the sets of stems across languages (i.e., English is our "tertium comparationis" for the comparison of lexical richness across languages). These translation equivalencies (TEs) are retrieved using the method of Liu et al. (2023), which iteratively extracts a substring (treated as the stem) from a target language that co-occurs most significantly with the seed item.[1]

Given this validation, we expect that using English as a medium for extracting colexifications should not pose problem as the mappings would be reversible and using other languages as a medium should yield equivalent topologies. All stems that occur $N = 1$ are omitted from subsequent analysis to avoid a long tail of overwhelmingly erroneous extractions. Tables 2 and 3 illustrates the retrieval step: given English 'land', the procedure extracts first the stem *wha:s* for Beja, with word forms *wha:s, wha:si* and *wha:so*, then the stem *tigadi:ri*, with only one word form, and so forth. In total, 6 stems with 9 word forms are extracted for Beja given the meaning (seed item) 'land'.

Colexification: To assess if sets of meanings (seed items) are frequently colexified, we first determine for individual languages if they colexify multiple meanings. The translation equivalencies between seed items and specific word forms in the target language are used to compute the **colexification matrix** for each language. Let $\mathbf{D}_x \in 0, 1^{|S|\times|V_t|}$ be a matrix with $|S|$ rows (the number of seed items) and $|V_t|$ columns (the number of unique specific word forms in the target language t). The colexification matrix $\mathbf{M}_t \in 0, 1^{|S|\times|S|}$ is then given as: $\mathbf{M}_t = \min(\mathbf{D}_t \cdot \mathbf{D}_t^T, 1)$. Each cell in $\mathbf{M}_t$ then represents whether a pair of shared seed items are colexified by at least one word form in a given language t. Table 4 shows how colexifications are extracted from the TEs in Table 2 into a colexification matrix.

While this procedure reliably captures various true colexifications (e.g., 'land', 'earth', and 'country' in Beja *wha:s*), it also captures pairs of meanings that are **synlexified** (Haspelmath, 2023). Synlexification involves the conflation of two concepts in a single word at the same time. For example,

[1] When evaluated on the match to the interlinear glosses provided for 26 of the languages, this method yields a highly reliable performance on a lexicon-wide sample for the DoReCo corpus, outperforming models based on word alignment (Beekhuizen, 2025).

'fall' and 'asleep' are often extracted as colexified in our procedure because many languages have a stem and hence word forms meaning 'to fall asleep' (e.g., *addormentarsi* in Italian). These are then the translation equivalent both of 'fall' and of 'asleep'. These synlexifications would introduce noise in the detection of lexical fields. A simple heuristic to remove them relies on the fact that for synlexification both (colexified) seed items are frequently present in the same (English) sentence. Letting S_a and S_b be the sets of sentences in which seed items a and b appear in the corpus of a given language t, we remove any pair a and b for which $S_a \cap S_b \neq \emptyset$.

Next, the languages' colexification matrices are summed to obtain a **crosslinguistic colexification matrix** representing how many languages colexify each pair of meanings. This colexification matrix is computed as $\mathbf{C} = \sum_t \mathbf{M}_t$, so that $\mathbf{C} \in \mathbb{Z}_{>0}^{|S|\times|S|}$. Table 5 shows an excerpt for the meanings 'earth', 'country', 'land', 'world' and 'ground'.

To derive lexical fields that are reliable cross-linguistically, we want to retain colexifications that are represented in a multiple languages. Therefore, we set to zero any pair of seed items that are colexified only once across languages.

Field Detection: If we consider this connectivity matrix as a graph, many of the concepts are connected; the largest connected component includes 211 seed items. To have manageable sets of seed items for which we can calculate the lexical richness, we have to carry out some form of finer-grained community detection on the graph. To that end, we first run k-clique percolation (Palla et al., 2005), with $k = 3$ to determine any coherent larger fields. This ignores pairs of recurrently colexified concepts (as they are cliques with $k = 2$), and as such we furthermore add each pair of colexified seed items that is not already a subgraph of any detected community. This creates 265 lexical fields. To present some examples of fields extracted, we obtained {'beautiful', 'nice', 'good'} and {'path', 'way', 'road'}, alongside fields capturing general abstract concepts of MORE or LESS: {'large', 'big', 'old', 'many', 'elder'} and {'little', 'small', 'young', 'few', 'son'}. Figure 1 represents a field, indeed corresponding to the five meanings in Tables 2 and 3, that connects the concrete concepts of 'land' and 'ground' with the abstract concepts of 'country' and 'world'.

Merging: As a final step, the stems (per languages) are merged using the method of Beekhuizen (2025). As the stems are determined per meaning, they are not guaranteed to be identical across meanings in a field. As such, this step overcomes this issue by merging stems with low edit distances.

Analysis

For the statistical analysis, we aim to model two measures of lexical richness in a target language t given a field F. First, as a naive measure we take the log-transformed **count** N of merged stems that a language t associates with the meanings in a field F, noted $N(F,t)$: a language with more terms given a field can be thought of a having a greater lexical richness. Second, as a measure that reflects the variable frequency of

Beja	beja1238	$f_{\text{narr}} = 1147$, $f_{\text{non-narr}} = 4635$, $N = 6$, $H = 1.23$
Seed	Stem	Form
land	wha:s [4]	*wha:s, wha:si, wha:so:*
	tigadi:ri [2]	*tigadi:ri*
	ist? [2]	*ist?e:, ist?ije:t*
	gwida:tu:jt [1]	*gwida:tu:jt*
	w?ardo: [1]	*w?ardo:*
	dir?a:t [1]	*dir?a:t*
ground	fitik'ti:t [1]	*fitik'ti:t*
	itirifif [1]	*itirifif*
	malja: [1]	*malja:*
earth	ha:s [6]	*ha:s, wha:s, w'ha:sida, ha:si:b*
	idibba [1]	*idibba*
country	ha:s [20]	*wha:s, ha:s, wha:si:b, wha:si, ha:sib, ha:si:b, wha:swwa,o:ngwha:sijo:n, tiwha:sib, wha:sija*
	bal [4]	*balad, balat, baldi:m*
world	tidijj [3]	*tidijja:ti:b, tidijja:tib*

Table 2: Translation Equivalencies of Beja for 'land', 'ground', 'earth', 'country', and 'world'. Narrative frequency f_{narr} (per million), non-narrative frequency $f_{\text{non-narr}}$, stem count N and entropy H of stem distribution given the field are indicated. The corpus frequency of each stem is indicated in brackets, with grey stems removed as they have $N = 1$. The colexification of 'land', 'earth', and 'country' is shown in red.

Hoocąk	hoch1243	$f_{\text{narr}} = 852$, $f_{\text{non-narr}} = 1498$, $N = 3$, $H = 1.10$
Seed	Stem	Form
land	maa [2]	*maa*
	hakikawa?uwi [1]	*hakika(w)a?uwi*
ground	maija [2]	*maija, maijasanasge*
earth	hake [2]	*hake*
country	wawok'uire [1]	*wawok'uire*
world	zeejakjene [1]	*zeejakjene*

Table 3: Translation Equivalencies for Hoocąk for 'land', 'ground', 'earth', 'country', and 'world' presenting information analoguous to Table 2

Beja (beja1238)	earth	country	place	land
earth	0	1	0	1
country	1	0	0	1
place	0	0	0	0
land	1	1	0	0

Table 4: Excerpt from Beja's colexification matrix. Cells indicate whether two concepts are colexified (1) or not (0).

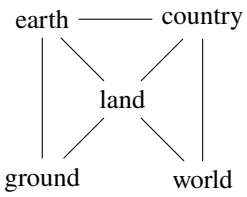

Figure 1: Graph of the relationships among the five concepts in the crosslinguistic colexification matrix of Table 5.

	earth	country	land	world	ground
earth	0	3	4	0	3
country	3	0	4	2	0
land	4	4	0	2	6
world	0	2	2	0	0
ground	3	0	6	0	0

Table 5: Excerpt of the crosslinguistc colexification matrix. Cells indicate how many languages colexify a pair of concepts.

the merged stems associated with the seed items in F, we use the **entropy** H over their frequency distribution, noted $H(S_t|F,t)$, where S_t is the set of merged stems for t. Entropy is more conservative, as it reflects not just the number of merged stems, but also their relative frequency: a field with four terms in a [.70, .10, .10, .10] distribution is arguably less rich than one with a [.25, .25, .25, .25] distribution.

The dependent variable is the count or entropy measure per field per language. As predictors, we use two measures of usage frequency of the field in that language, both defined as the log-transformed frequency-per-million (log FPM) of each field in a subcorpus of the language. The log FPM in the narratives for a language t is given as f_{narr}, and the log FPM in non-narrative genres as $f_{\text{non-narr}}$. We remove from consideration any cases where, for a given language, a lexical field does not appear in both the narrative and non-narrative subcorpora. We then carry out model comparison over six mixed-effect linear regression models: two dependent variables by three variants of the independent variables: only f_{narr}, only $f_{\text{non-narr}}$, and both f_{narr} + $f_{\text{non-narr}}$ All models have two random effects: the language t and the field F, included to absorb variance associated with fields or languages.

Since the corpora of different languages are of different sizes, a given field is not always fully represented in each corpus, which may lead to a poor estimate of lexical richness. For instance, a field of 3 seed items might only have TEs for 1 seed item in a given language's corpus, making it impossible to estimate how that language would express the other two seed items. Thus, we compute the saturation of a field F in a given language x as: $sat_x(F) = \frac{|F \cap T_x|}{|F|}$ where T_x represents the translation equivalencies of x. For a given field, we eliminate languages that do not fully saturate it, resulting in 3200 observations across 26 languages. (Given small corpus sizes, many languages are eliminated, as they lack fields that occur in both subcorpora and that are fully saturated.)

Results

Model comparison using Bayes Information Criterium (BIC), presented in Table 6, shows that both models based on narrative frequency explain the measures of lexical richness better than the non-narrative models. Considering both frequency estimates achieves slightly greater explanatory power, despite the additional parameter, but the gap between Non-Narrative and Narrative is greater than between Narrative and the combined model. This supports the idea that narrative frequency models the communicative need better than the frequency in other genres, though the other genres do make an independent explanatory contribution.

A comparison of the strength of the effects of f_{narr} vs. $f_{\text{non-narr}}$ can be found in the Combined model, reported in Tables 7 (Count) and 8 (Entropy). In both cases, the effect of narrative frequency (Count: $\beta = 0.23$, Entropy: $\beta = 0.15$) is almost twice stronger than that of non-narrative frequency (Count: $\beta = 0.12$, Entropy: $\beta = 0.08$). A t-test between the estimated slopes for narrative and non-narrative frequency shows that this difference is significant: $t = 8.53; p \ll .001$ for the Count model, and $t = 4.18; p \ll .001$ for the Entropy model. Further of interest is the fact that the correlation between f_{narr} and $f_{\text{non-narr}}$ is $r = -0.37$ for the Count model, and $r = -0.40$ for the Entropy model: a weak-medium correlation, meaning that a field's frequencies in narrative and non-narrative data are mostly independent – the negative direction of the correlation is a curiosity for future research. Finally, Figure 2 shows the model fits and correlation between data and predictions in the narrative/non-narrative and count/entropy models.

	Count	Entropy
Non-Narrative	3651.7	5343.2
Narrative	3368.3	5268.4
Narrative and Non-Narrative	3173.8	5233.2

Table 6: BIC for the three models per dependent variable.

Fixed Effects	β	SE	t-value	p-value
(Intercept)	-0.89	0.10	-9.50	$\ll$.001
f_{narr}	0.23	0.01	24.09	$\ll$.001
$f_{\text{non-narr}}$	0.12	0.01	14.67	$\ll$.001
Random Effects: field (VE=0.07), language (VE=0.08)				
3200 Observations; 260 Fields; 26 Languages				

Table 7: **Count** Regression model (VE=variance explained).

Our results show that, across the languages analyzed, a lexical field with a higher narrative frequency is associated with a higher number of merged lemmas and a higher entropy over their frequency distribution. The examples in Tables 2-3 demonstrate this principle. Beja has a narrative frequency $f_{\text{narr}} = 1147$ (per million) with $N = 6$ merged lemmas and an entropy of $H = 1.23$ over them. In comparison, Hoocąk has a

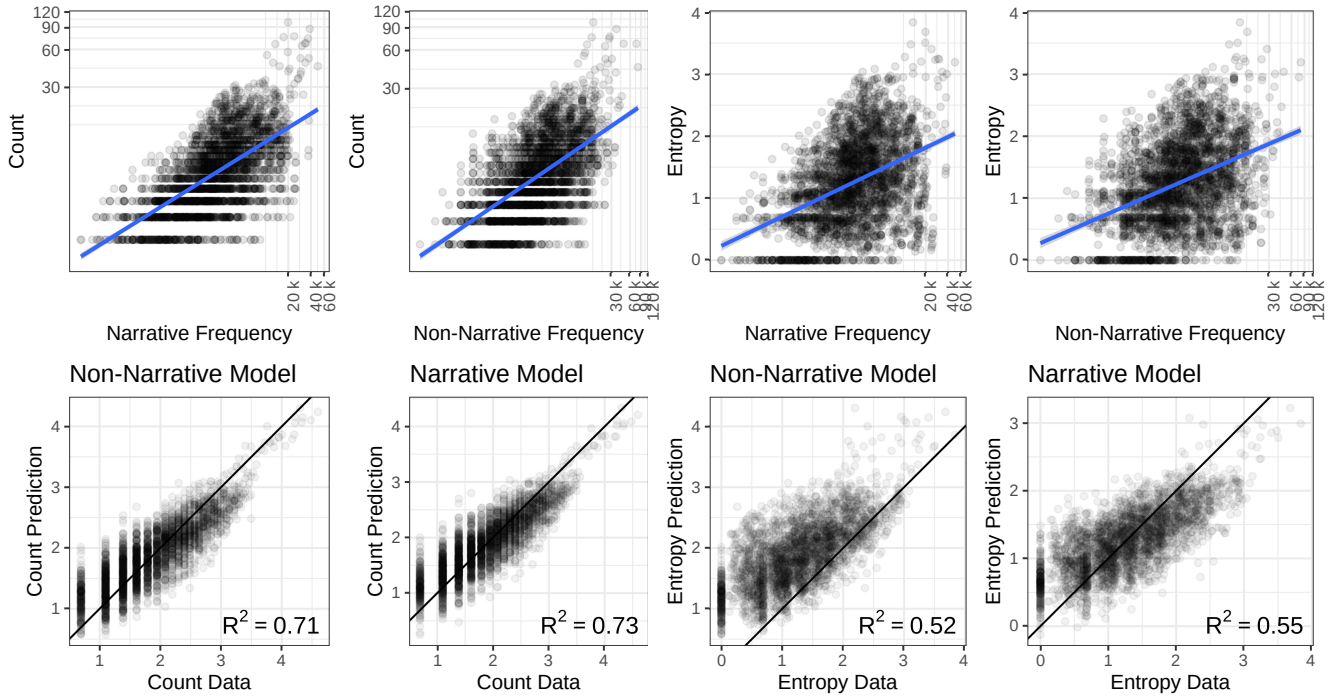

Figure 2: Comparison between Narrative and Non-Narrative models. Upper row: model fit. Lower row: correlation between dependent variables and predictions.

Fixed Effects	β	SE	t-value	p-value
(Intercept)	-0.57	0.11	-5.08	$\ll$.001
f_{narr}	0.15	0.01	11.51	$\ll$.001
$f_{\text{non-narr}}$	0.08	0.01	6.71	$\ll$.001
Random Effects: field (var.=0.09), language (var.=0.07)				
3200 Observations; 260 Fields; 26 Languages				

Table 8: **Entropy** Regression model (VE=variance explained)

lower narrative frequency $f_{\text{narr}} = 852$ (per million) with only $N = 3$ merged lemmas and a lower entropy of $H = 1.10$.

Discussion

In this paper, we present a novel finding: usage frequency in narrative, more so than in non-narrative genres, explains a language's lexical richness. This finding is explained, we argue, by communicative properties of narrative that make a semantically fine-grained lexicon adaptive. Our results show that the richness of lexical fields – measured as the number of unique terms expressing the field or the uncertainty over those terms – is better explained by the frequency of usage of the field in narrative contexts than its frequency in a non-narrative contexts. In future work, we intend to further substantiate this finding by considering more conservative methodologies and permutations of the present methodological choices. Similarly, we aim to consider potential alternative explanations of the result and control for them where possible. Most importantly, further theoretical work is necessary to place this finding within (formal) accounts of the lexicon and provide a more mechanistic explanation. We suggest that this could be accomplished by having a variable "penalty" on the semantic error, that is: how much the recipient's construal of the situation deviates from the speaker's intended construal. Narrative might then have a higher penalty than joint deliberation or instruction, thus leading to a greater importance of informativity in the language when a domain is frequent.

Our findings show that communicative need is not a monolithic concept. The informativity of a given semantic system is not only motivated by usage in general, but appears weighted by the varying pressures that different pragmatic goals may bring. As such, they present a new direction of empirical exploration within the efficient communication framework, that aims to understand how different communicative contexts may all affect a speaker's desire to be informative differently.

This direction of research would suggest a further reorientation to how we consider the relation between language as a (mental, social) system and the contexts in which language is used. Rather than accidental places where language happens to be deployed, these embeddings can be argued to be intimately interwoven with the language that is used in them, with neither truly existing without the other (Bruner, 1990; Duranti & Goodwin, 1992; Hanks, 1996). Such a reorientation would potentially open the door for deeper explanations of why lexical semantic inventories are the way they are.

References

Anand, G., & Regier, T. (2023). Kinship terminologies reflect culture-specific communicative need: Evidence from hindi and english. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Aznar, J. (2024). Nisvai DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/nisv1234 doi: 10.34847/nkl.2801565f

Bakhtin, M. (1986). *Speech genres & other late essays*. Austin, TX: University of Texas Press.

Bartels, H., & Szczepański, M. (2024). Lower Sorbian DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/lowe1385 doi: 10.34847/nkl.6c6e4e9k

Beekhuizen, B. (2024). Discursive distinctiveness explains lexical differences between languages. In J. Nolle et al. (Eds.), *The Evolution of Language: Proceedings of the 15th International Conference (Evolang XV).* Retrieved from https://evolang2024.github.io/proceedings/paper.html?nr=147 doi: 10.17617/2.3587960

Beekhuizen, B. (2025). Token-level semantic typology without a massively parallel corpus. In *The 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Bietti, L. M., Tilston, O., & Bangerter, A. (2019). Storytelling as adaptive collective sensemaking. *Topics in cognitive science*, *11*(4), 710–732.

Bogomolova, N., Ganenkov, D., & Schiborr, N. N. (2024). Tabasaran DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/taba1259 doi: 10.34847/nkl.ad7f97xr

Boyd, B. (2018). The evolution of stories: From mimesis to language, from fact to fiction. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(1), e1444.

Bradford, L., Thomas, G., & Xu, Y. (2022). Communicative need modulates lexical precision across semantic domains: A domain-level account of efficient communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Bruner, J. (1990). *Acts of meaning: Four lectures on mind and culture* (Vol. 3). Harvard university press.

Burenhult, N. (2024). Jahai DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/jeha1242 doi: 10.34847/nkl.6a71xp0p

Cobbinah, A. Y. (2024). Baïnounk Gubëeher DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/bain1259 doi: 10.34847/nkl.a332abw8

Corballis, M. C. (2009). Mental time travel and the shaping of language. *Experimental Brain Research*, *192*(3), 553–560.

Cowell, A. (2024). Arapaho DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/arap1274 doi: 10.34847/nkl.36f5r1b6

Cuff, E. C. (1993). *Problems of versions in everyday situations*. University Press of America.

Duranti, A., & Goodwin, C. (1992). *Rethinking context: Language as an interactive phenomenon* (No. 11). Cambridge University Press.

Däbritz, C. L., Kudryakova, N., Stapert, E., & Arkhipov, A. (2024). Dolgan DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/dolg1241 doi: 10.34847/nkl.f09eikq3

Döhler, C. (2024). Komnzo DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/komn1238 doi: 10.34847/nkl.c5e6dudv

Everett, D. (2010). *Don't sleep, there are snakes: Life and language in the amazonian jungle*. Profile books.

Ferretti, F., & Adornetti, I. (2020). Why we need a narrative brain to account for the origin of language. *Paradigmi*, *38*(2), 269–292.

Forker, D., & Schiborr, N. N. (2024). Sanzhi Dargwa DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/sanz1248 doi: 10.34847/nkl.81934177

François, A. (2008). Semantic maps an the typology of colexifications: Intertwining polysemous networks across languages. In M. Vanhove (Ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations* (pp. 163–216). Amsterdam: John Benjamins.

Franjieh, M. (2024). Fanbyak DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/orko1234 doi: 10.34847/nkl.02084446

Garcia-Laguia, A. (2024). Northern Alta DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/nort2875 doi: 10.34847/nkl.efea0b36

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790.

Gipper, S., & Ballivián Torrico, J. (2024). Yurakaré DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/yura1255 doi: 10.34847/nkl.7ca412wg

Gippert, J. (2024). Svan DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/svan1243 doi: 10.34847/nkl.9ba054c3

Griscom, R. (2024). Asimjeeg Datooga DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/tsim1256 doi: 10.34847/nkl.f77c7m72

Gusev, V., Klooster, T., Wagner-Nagy, B., & Arkhipov, A. (2024). Kamas DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/kama1351 doi: 10.34847/nkl.cdd8177b

Güldemann, T., Ernszt, M., Siegmund, S., & Witzlack-Makarevich, A. (2024). Nng DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/nngg1234 doi: 10.34847/nkl.f6c37fi0

Haig, G., Vollmer, M., & Thiele, H. (2024). Northern Kurdish (Kurmanji) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/nort2641 doi: 10.34847/nkl.ca10ez5t

Hanks, W. F. (1996). *Language & communicative practices*. Routledge.

Hartmann, I. (2024). Hoocąk DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/hoch1243 doi: 10.34847/nkl.b57f5065

Harvey, A. (2024). Gorwaa DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/goro1270 doi: 10.34847/nkl.a4b4ijj2

Haspelmath, M. (2018). How comparative concepts and descriptive linguistic categories are different. In D. Olmen, T. Mortelmans, & F. Brisard (Eds.), *Aspects of linguistic variation* (pp. 83–114). De Gruyter Mouton.

Haspelmath, M. (2023). Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, *14*, 1236853. Retrieved from https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1236853/full doi: https://doi.org/10.3389/fpsyg.2023.1236853

Haude, K. (2024). Movima DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/movi1243 doi: 10.34847/nkl.da42xf67

Hellwig, B. (2024). Goemai DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/goem1240 doi: 10.34847/nkl.b93664ml

Hellwig, B., Schneider-Blum, G., & Ismail, K. B. K. (2024). Tabaq (Karko) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/kark1256 doi: 10.34847/nkl.eea8144j

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Kazakevich, O., & Klyachko, E. (2024). Evenki DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/even1259 doi: 10.34847/nkl.5e0d27cu

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).

Kim, S.-U. (2024). Jejuan DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/jeju1234 doi: 10.34847/nkl.06ebrk38

Krifka, M. (2024). Daakie DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/port1286 doi: 10.34847/nkl.efeav5l9

Labov, W. (2013). *The language of life and death: The transformation of experience in oral narrative*. Cambridge University Press.

Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In J. Helm (Ed.), *Essays on the Verbal and Visual Arts. Proceedings of the 1966 Spring Meeting of the American Ethnological Society* (pp. 12–44).

Langacker, R. W. (2019). Construal. In E. Dąbrowska & D. Divjak (Eds.), *Cognitive linguistics: Foundations of language* (pp. 140–166). Berlin: Mouton de Gruyter.

Levinson, S. C. (1979). Activity types and language. *Linguistics*, *17*, 365–399.

Liu, Y., Ye, H., Weissweiler, L., Wicke, P., Pei, R., Zangenfeind, R., & Schütze, H. (2023). A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 12969–13000).

Michaud, A. (2024). Yongning Na DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/yong1270 doi: 10.34847/nkl.abe65p95

Mosel, U. (2024). Teop DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/teop1238 doi: 10.34847/nkl.9322sdf2

Ozerov, P. (2024). Anal DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/anal1239 doi: 10.34847/nkl.0dbazp8m

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *nature*, *435*(7043), 814–818.

Ponsonnet, M. (2024). Dalabon DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/ngal1292 doi: 10.34847/nkl.fae299ug

Quesada, J. D., Skopeteas, S., Pasamonik, C., Brokmann, C., & Fischer, F. (2024). Cabécar DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/cabe1245 doi: 10.34847/nkl.ebc4ra22

Reiter, S. (2024). Cashinahua DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/cash1254 doi: 10.34847/nkl.a8f9q2f1

Riesberg, S. (n.d.). dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Ring, H. (2024). Pnar DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/pnar1238 doi: 10.34847/nkl.5ba1062k

Rose, F. (2024). Mojeño Trinitario DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/trin1278 doi: 10.34847/nkl.cbc3b4xr

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., ... others (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, *7*(1), 13.

Schiborr, N. N. (2024). English (Southern England) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/sout3282 doi: 10.34847/nkl.9c271u5g

Schiffrin, D. (1996). Narrative as self-portrait: Sociolinguistic constructions of identity. *Language in society*, *25*(2), 167–203.

Schnell, S. (2024). Vera'a DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/vera1241 doi: 10.34847/nkl.3e2cu8c4

Seifart, F. (2024a). Bora DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/bora1263 doi: 10.34847/nkl.6eaf5laq

Seifart, F. (2024b). Resígaro DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/resi1247 doi: 10.34847/nkl.ffb96lo8

Seifart, F., Paschen, L., & Stave, M. (2024). *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr doi: 10.34847/nkl.7cbfq779

Skopeteas, S. (2024). Yucatec Maya DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/yuca1254 doi: 10.34847/nkl.9cbb3619

Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J., & Schröter, S. (2024). Urum DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/urum1249 doi: 10.34847/nkl.ac166n10

Swales, J. M. (1990). *Genre analysis*. Cambridge university press.

Teo, A. (2024). Sümi DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/sumi1235 doi: 10.34847/nkl.5ad4t01p

Thieberger, N. (2024). Nafsan (South Efate) DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/sout2856 doi: 10.34847/nkl.ba4f760l

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Vanhove, M. (2024). Beja DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/beja1238 doi: 10.34847/nkl.edd011t1

Vydrina, A. (2024). Kakabe DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/kaka1265 doi: 10.34847/nkl.d5aeu9t6

Wegener, C. (2024). Savosavo DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/savo1255 doi: 10.34847/nkl.b74d1b33

Wichmann, S. (2024). Texistepec Popoluca DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/texi1237 doi: 10.34847/nkl.c50ck58f

Witzlack-Makarevich, A., Namyalo, S., Kiriggwajjo, A., & Molochieva, Z. (2024). Ruuli DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/ruul1235 doi: 10.34847/nkl.fde4pp1u

Xu, X., & Bai, B. (2024). Sadu DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Retrieved 12/08/2024, from https://doreco.huma-num.fr/languages/sadu1234 doi: 10.34847/nkl.3db4u59d

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2020). Communicative need in colour naming. *Cognitive Neuropsychology*, *37*(5-6), 312–324.