Syntacticization Across Registers, Genres and Modalities: A Crosslinguistic Quantitative Cartographic Study

Giuseppe Samo (Giuseppe.Samo@unige.ch; samo@blcu.edu.cn)

1. Introduction

The Left Periphery of the clause (LP) and/or the Complementizer Phrase (CP) has represented an excellent locus of investigation in syntactic cartography (Rizzi 1997, Puskás 2000, Aboh 2004, Benincà & Poletto 2004, Ledgeway 2010, Wolfe 2015 *inter alia*) especially in terms of syntacticization of scope-discourse properties (Cinque & Rizzi 2010, Si 2011, Rizzi & Cinque 2016). Take as an example the two sentences in Italian given in (1).

(1) a. Luigi ha scritto questo articolo nel 1997.
Luigi has written this article in 1997
‘Luigi has written this article in 1997.’
b. Quest’articolo, Luigi l’ha scritto nel 1997.
this article Luigi it-has written in 1997
‘This article, Luigi wrote in 1997.’

The two sentences differ in terms of topicality. In (1a), the topicality—assuming no prosodic cues (Bocci 2013)—concerns the entire clause, whereas in (1b), the preposed internal argument of the verb, *questo articolo* (‘this article’), is topicalized. The sentences also differ in the contexts in which they can be used: for instance, as noted by Rizzi (2004), while (1a) can serve as an answer to a “what happened?” question, (1b) cannot.

Topics, or any constituent bearing discourse properties, are closely related to information transfer and, therefore, to extra-linguistic factors. In this sense, grammars are sensitive to variation across registers or genres. For example, several cartographic studies have shown that non-pro-drop languages like English may permit null subjects in certain contexts, such in diary-style registers (Haegeman 1997, 2013; see also Endo 2022 for question formation in comics). Also, it has been noted that structures involving the fronting of objects in the Left Periphery—are reduced when the sentences represent the content of emergency knowledge-transfer in medical contexts (Zhao et al. 2021 and related works).

Given the existence of large-scale databases—whether heterogeneric (spanning multiple genres) or monogeneric (limited to a single genre)—is it possible to quantify and classify registers from a syntactic perspective, for example, by examining the activation of the CP layer through object fronting?

To answer these questions, we conduct a study in Quantitative Cartography (Samo 2019b, 2022, 2023, 2025, Samo & Isolani 2024), adopting the guiding principles of Quantitative Computational Syntax (Merlo 2016 and related works), in which large-scale datasets and simple computational models are explored to test the predictions of linguistic proposals.

To achieve our goal, the paper is structured as follows. Section 2 outlines the core elements of sentences in which the LP is activated and discusses how a theory of genres and registers can be integrated, along with the cross-linguistic dimension of the syntactically annotated corpora we examine. Section 3 presents the study and its results. Section 4 provides a discussion of the findings. Finally, Section 5 concludes the paper.

2. BACKGROUND

The syntactic derivations of the sentences in (1) are given in (2), where EA stands for ‘external argument’ and IA ‘internal argument’.

(2) a. [$_{\text{SpecCP}}$ [$_{\text{C°}}$ [$_{\text{SpecIP}}$ Luigi [$_{\text{I°}}$ scrisse [$_{\text{EA}}$ <Luigi> [$_{\text{v°}}$ <scrisse>[$_{\text{IA°}}$ questo libro]]]]]]]
b. [$_{\text{SpecCP}}$ Questo libro [$_{\text{C°}}$ [$_{\text{SpecIP}}$ Luigi [$_{\text{I°}}$ lo scrisse [$_{\text{EA}}$ <Luigi> [$_{\text{v°}}$ <scrisse> [$_{\text{IA°}}$ <questo libro> <lo>]]]]]]]

While (1a) represents an instance of a canonical clause, in which the highest activated portion of the syntactic architecture is the Specifier of the IP - the position dedicated to subjects (Cardinaletti 2004), the sentence in (1b) shows the activation of the CP layer.[1] The activation of the CP layer contributes to the understanding of scope-discourse properties, understood as “the scope of operators and the expression of discourse-related properties linked to the informational organization of the sentence, such as topicality and focus.” (Rizzi 2014:517). Their activation leads to reorderings, which are closely related to syntactic complexity in parsing.

The study of syntactic complexity in Cartography—understood as deviation from canonical clause structure through movement and reordering —has long been central to both theoretical linguistics and experimental research in acquisition, developmental and adult grammars (Friedmann et al. 2009, Durrleman et al. 2016, Rizzi 2016). However, complexity does not arise solely from syntactic structure; it is also shaped by extra-syntactic factors such as genre, register, and communicative intent. In particular, corpus linguistics has highlighted the importance of genre classification and the cartographic mapping of syntactic variation across text types (Bybee & Thompson 2021 *inter alia*). Syntactic complexity can be defined computationally and applied across genres and registers in corpus-based studies.

A generative/cartographic approach has been explored in domains like English null-subjects in diary writing (Haegeman 1997, 2013) and comics (Endo 2022). The focus of this type of research is that a parameter value that leads to ungrammaticality, like the null subject parameter in English (Rizzi 1982) can be produced in very specific contexts, like diary contexts. Another genre that gained investigation in syntactic cartography has been medical-content communication in emergency situations. As Gamhewage et al. (2020) observed, effective knowledge transfer during emergencies must not only provide accurate information but also reduce the cognitive effort required to process it. In this context, emergency communication tends to strategically minimize linguistic complexity—particularly regarding syntactic locality. For instance, Zhao et al. (2021) found that complex A′-configurations were significantly reduced in emergency health texts produced by OpenWHO for COVID-19 responders, especially when compared to corpora drawn from social media, news, or legal documents.

However, large-scale datasets often use broad genre labels that lack the granularity necessary for fine-grained linguistic analysis. While these labels are useful for macro-level categorization, they often fall short when precision and domain-specific nuance are required. In addition, the syntactic relation labels used in these datasets are typically broad, further limiting their usefulness for fine-grained linguistic inquiry. As a result, a detailed typology of

[1] The sentences in (1b) and (2b) can also be derived via base-generation of the left-peripheral Topic (Wolfe 2022). Whether the Topic is moved or base-generated, the results of our study would remain unchanged.

cartographic labels cannot be derived semi-automatically from existing datasets. Building on the methodology outlined in Samo (2019b), we propose translating accessible data from existing sources into cartographic representations as a way to test cartographic hypotheses. Our approach uses a limited set of syntactic and morpho-syntactic labels based on the Universal Dependencies framework (Nivre 2015, de Marneffe et al. 2021), which facilitates cross-linguistic comparison, since hundreds of treebanks across languages are (morpho)-syntactically annotated using the same annotation schema.

Before proceeding to the study, we present some notes on the genres that we discuss. Encyclopedic entries typically represent a formal register characterized by a high degree of lexical precision. Their purpose is to convey factual, possibly neutral and structured knowledge. A significant subset of these resources is on the web—such as *Wikipedia* or *Baidu Baike* (百度百科)—and are freely available, but also they can be community-curated platforms. They can be created and maintained on a volunteer basis (Margaretha & Lüngen 2014, Ursini & Samo 2025, and references therein), and offer extensive multilingual coverage, making them valuable for cross-linguistic studies (cf. Poudat et al. 2024). They also include dialects and under-resourced languages. Publicly available sources like Wikipedia have also served as training data for crosslinguistic neural language models (Gulordava et al. 2018), further attesting to their centrality in the late 2010s in both linguistic and computational research. Social media texts (e.g., tweets, posts, comments) represent a highly interactive, informal, and often multimodal register, with the frequent use of images, hashtags, emojis, and other paralinguistic markers (see also Daniel & Camp 2020). They often mix informational and opinionative content and display wide variation depending on platform and user intents (cf. Androutsopoulos 2014).

Spoken modality and texts, such as transcribed conversations, interviews, or oral narratives, constitute a “register” marked by spontaneous production and frequent use of discourse markers, hesitations, repetitions, and incomplete structures. These features reflect the interactional nature of speech and its reliance on shared context and co-presence. From a corpus perspective, spoken data is central to understanding everyday language use and variation (Biber et al. 1999). Partially related to spoken text, we can find fictional narrative, which often blends narrative and dialogic passages, which allows it to reflect both written and spoken-like features, even in contexts such as science-fiction (Rüdiger & Lange 2023). News reports form a journalistic register (Bell 1991) that tends to favor concise, clear syntax, a high proportion of proper nouns, and formulaic expressions, particularly in headlines. Domain-specific texts can likewise be investigated as registers in their own right: legal, but also medical, discourse exemplifies a highly formal, specialized register characterized by syntactic complexity, fixed phraseology (e.g., *legalese* for Legal*,* Martínez et al. 2023).

However, most of the treebanks that we will study (see details in Section 3) are heterogeneric, comprising texts drawn from multiple genres and registers. This diversity can complicate linguistic analyses, as syntactic patterns may vary significantly depending on the communicative context, formality level, or modality of the source material. Finally, one type of heterogeneric corpus that we will analyze is derived from parallel treebanks (i.e., the same content in translation; cf. Ahrenberg 2007, Volk et al. 2014), which facilitate cross-linguistic comparison. In our case, we analyze, for example, the Parallel Universal Dependencies treebank, which consists of news and Wikipedia texts (see details in Section 3).

Section 3 presents the study, the underlying hypotheses, materials, methodology, while Section 4 discusses the results.

3. MATERIALS & METHODS

Our study takes a more observational approach, aiming to determine whether there is a cross-linguistic correlation between object fronting, as in (2b), and specific genres. To evaluate this, we adopt a straightforward computational metric from Samo et al. (2020): the ratio of instances in which the object is fronted to those in which the internal argument of the verbal domain remains in its canonical post-verbal position. A higher ratio indicates greater syntactic flexibility within a genre, corresponding to increased activation of the Left Periphery. That is, at least one pair of genres shows a significantly different ratio of object fronting, implying structural variation across genres. The alternative hypothesis is tested against a null hypothesis which assumes no correlation between the ratio and genre type.

As discussed in Sections 1 and 2, we examine 53 syntactically annotated treebanks from the Universal Dependencies (UD) project, covering 15 languages, as shown in Table 1. We only select SVO languages in main clauses (Dutch is the only SOV language, but V2 in root contexts). To support cross-linguistic comparison, we also add parallel treebanks available in the Parallel Universal Dependencies (PUD, version 2.15) collection. Each PUD treebank contains the same 1,000 sentences, extracted from journalistic and encyclopedic texts and presented in the same order across languages, though the number of tokens varies. Of these 1,000 sentences, 750 sentences originate in English, while the remaining 250 come from original texts in German, French, Italian, or Spanish. As noted by the treebank creators translations into the other languages were mediated *via* English.

Let us introduce the languages: for Arabic, we analyze the heterogeneric sources of PADT (Smrž et al. 2008), NYUAD and the Arabic PUD. For Chinese, we include GSDSimp, built on text in simplified characters of encyclopedic entries, HK (a treebank of subtitles and legislative texts in traditional Chinese, Leung et al. 2016), PatentChar, containing texts related to patent applications, and the Chinese PUD. The Czech data include the heterogeneric sources of CAC (Hladká et al. 2007), the Czech PUD, and the domain-specific treebanks of FicTree (literary texts; Jelínek 2017), CLTT (legal texts; Kríž & Hladká 2018). For Dutch, we explored the treebank LassySmall (van Noord et al. 2013), built from Wikipedia, and Alpino (Bouma & van Noord 2017), which is based on news. For English, we use the social media source GUMReddit (Behzad & Zeldes 2020), a repository that only contains annotations, without the underlying textual data from Reddit; the domain specific ATIS (Airline Travel Information System) dataset which includes the human speech transcriptions of people asking for flight information on the automated inquiry systems, the heterogeneric sources GUM (Zeldes 2017), EWT (Silveira et al. 2014) based on Web English, and LinES (Ahrenberg, 2015, which contains literature, an online manual, and Europarl data) and the PUD. For Estonian, we analyzed the EDT treebank (Muischnek et al. 2014) and a web treebank composed of blogs, social media and web (Muischnek et al. 2019). Beyond French PUD, French data are also retrieved from two heterogeneous written sources (GSD, Guillaume et al. 2019, Sequoia, Candito et al. 2014), and two oral corpora: ParisStories (Kahane et al. 2021) and Rhapsodie. For Hebrew we study two corpora, the news treebank HTB (McDonald et al. 2013) and the IATHLTWiki (Zeldes et al. 2022) containing encyclopedic entries. Irish data are extracted from the heterogeneous source IDT and the social media treebank TwittIrish (Cassidy et al. 2022). For Italian, we explored three heterogeneric corpora (ISDT, Bosco et al. 2013; VIT, Alfieri & Tamburini 2016; ParTUT, Sanguinetti & Bosco 2014), two social media corpora (PosTWITA, Sanguinetti et al. 2018; TWITTIRO, Cignarella et al. 2018) and the Italian PUD. The Portuguese datasets are the heterogeneous sources GSD and PUD, a corpus of news (Bosque, Rademaker et al. 2017) and the social media data DANTEStocks (di Felippo et al. 2024). We queried one heterogeneous source for Romanian (RRT, Mititelu 2018) and the treebank SiMoNERO (Mititelu & Mitrofan 2020), containing scientific books, journal articles and blog posts related to the medical domain.

Finally, Spanish data are collected from the heterogeneous treebank GSD (X) and PUD, the news treebank AnCora (Taulé et al. 2008) and the spoken data for Rural Spanish COSER (Bonilla 2024). Size and the labels we adopt for this study are presented in Table 1.

Language	**Treebank$^{\text{Genre}}$**	**Condition**	**Size (tokens in K)**
Arabic	PADT$^{\text{N}}$	News	282K
	NYUAD$^{\text{N}}$	News	738K
	PUD$^{\text{N, W}}$	PUD	20K
Chinese	GSDSimp$^{\text{W}}$	Encyclopedic	123K
	HK$^{\text{S}}$	H	9K
	PatentChar$^{\text{L}}$	Legal	5K
	PUD$^{\text{N, W}}$	PUD	18K
Czech	CAC$^{\text{A, L, N, NF, R}}$	H	495K
	PUD$^{\text{N, W}}$	PUD	18K
	CLTT$^{\text{L}}$	Legal	36K
	FicTree$^{\text{F}}$	Literature	167K
Dutch	LassySmall$^{\text{W}}$	Encyclopedic	297K
	Alpino$^{\text{N}}$	News	208K
English	GUMReddit$^{\text{BL, SM}}$	Social Media	16K
	Atis$^{\text{N, NF}}$	Communication	61K
	GUM$^{\text{A, BL, M, F, L, N, NF, SM, S, WEB, W}}$	H	234K
	EWT$^{\text{BL, M, R, SM, WEB}}$	Web	254K
	LinES$^{\text{F, NF, S}}$	H	106K
	PUD$^{\text{N, W}}$	PUD	21K
Estonian	EDT$^{\text{A, F, N, NF}}$	H	438K
	EWT$^{\text{BL, SM, WEB}}$	Web	90K
French	GSD$^{\text{BL, N, R, W}}$	H	400K
	Rhapsodie$^{\text{S}}$	Spoken	44K
	ParisStories$^{\text{S}}$	Spoken	42K

	Sequoia$^{M, N, NF, W}$	H	70K
	PUD$^{N, W}$	PUD	24K
Hebrew	IAHLT^{W}	Encyclopedic	140K
	HTB^{N}	News	160K
Irish	IDT$^{F, L, N, WEB}$	H	115K
	TwittIrish^{SM}	Social Media	47K
Italian	VIT$^{N, NF}$	H	280K
	PoSTWITA^{SM}	Social Media	124K
	ISDT$^{L, N, W}$	H	298K
	ParTUT$^{L, N, W}$	H	55K
	PUD$^{N, W}$	PUD	23K
	TWITTIRO^{SM}	Social Media	29K
	ParlaMint^{L}	Parliament	20K
Portuguese	DANTEstocks^{SM}	Social Media	80K
	Bosque^{N}	News	227K
	GSD$^{BL, N}$	H	318K
	PUD$^{N, W}$	PUD	23K
Romanian	RRT$^{A, F, L, M, N, NF, W}$	H	218K
	SiMoNERo^{M}	Medical	146K
Russian	GSD^{W}	Encyclopedic	97K
	SynTagRus$^{F, N, NF}$	H	1517K
	Taiga$^{BL, F, N, P, SM, W}$	H	1758K
	PUD$^{N, W}$	PUD	19K
Slovenian	SSJ$^{F, N, NF}$	H	267K
	SST^{S}	Spoken	98K

Spanish	AnCora^{N}	News	568K
	GSD$^{BL, N, R, W}$	H	431K
	PUD$^{N, W}$	PUD	23K
	COSER^{S}	Spoken	8K

Table 1: Language, treebank, genres (A = academic, BL = Blog, F = Fiction, M = Mails, N = News, NF = Non-fiction, P = Poetry, R = Reviews, S = Spoken, W = Wiki, WEB = Web) and size in tokens (in thousands).

Our queries aim to retrieve the counts of objects that precede or follow the verb. To avoid issues related to flexibility of the pronominal nature of objects (Benincà 1995, Fontana 1997), we only retrieve data corresponding to maximal projections (XP). To achieve this, we target internal arguments identified by the dependencies *obj*, *iobj*, and *comp:obj*, which are governed by a verb in a main clause (*root*). The head of such a dependency should have a morpho-syntactic part-of-speech (*upos*) of either *NOUN* (nominal entities) or *PROPN* (proper nouns). For languages and treebanks that support it, we further restrict the search to finite verb forms (VerbForm=Fin). The queries are presented in Table 2, along with an example from Italian. All the data are queried and automatically retrieved from a python environment on count.grew.pl.[2]

Condition	Query	Example
V - Arg	pattern { X-[root]-> Verb; Verb -[obj | iobj | comp: obj]-> Arg; Verb << Arg; Arg [upos = "NOUN" | "PROPN"] (; VerbForm = Fin])}	*tutti possono usare* ***questa parola*** 'Everyone can use this word' (VIT, VIT-4012)
Arg - V	pattern { X-[root]-> Verb; Verb -[obj | iobj | comp: obj]-> Arg; Arg << Verb; Arg [upos = "NOUN" | "PROPN"] (; VerbForm = Fin])}	*Io* ***il privato*** *lo concepisco come un metodo di lavoro* 'I, the private, understand it as a working methodology' (VIT-217)

Table 2: Condition, queries and an example from Italian.

Optionality in general never matches with the goals of generative syntactic approaches (Samo & Si 2022). Following Samo & Merlo (2019), we conducted a binomial test on the raw numbers to compare the results within conditions and between treebanks, and to exclude outcomes that might be due to chance. Accordingly, we calculated a z-score and a p-value. A high absolute value of the z-score (whether positive or negative) can be interpreted as stronger evidence against the null hypothesis (i.e., no observable trends) for each condition. For each language, we calculate the z-score and the p-value on the largest treebank. Section 3.3. presents the results of the study.

[2] All queries and results are available at the following link: https://github.com/samo-g/arg-lp.git

3.3. Results & Discussion

As mentioned in Subsection 3.1, our metric is based on the ratio between fronted objects and objects in the canonical position. The higher this ratio, the stronger the observed trend toward activation of left peripheral positions for internal arguments. Figure 1 summarizes the results.

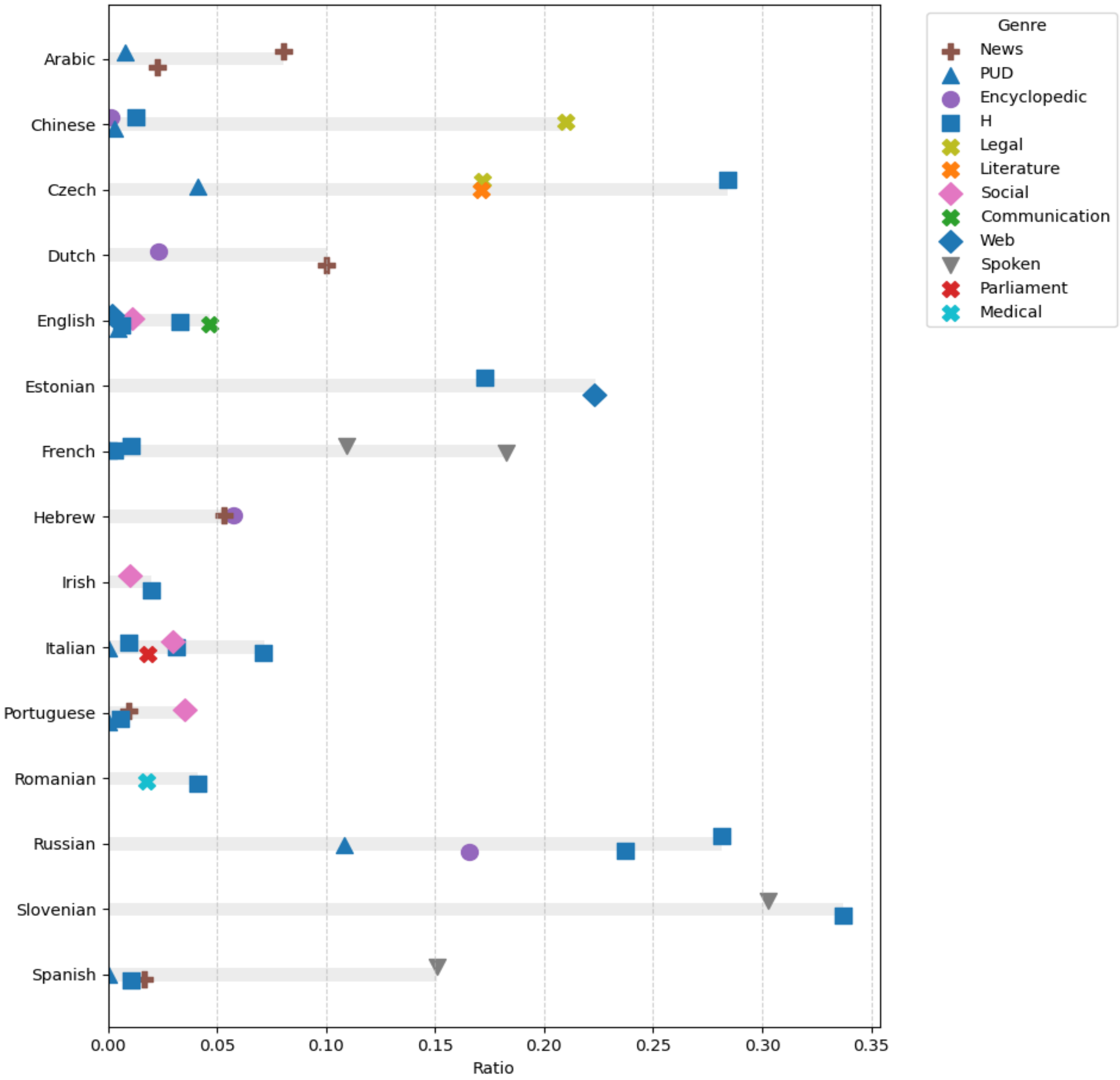

Figure 1.

As Figure 1 shows, Slavic languages (Haider & Szucsich 2022) show the highest ratios, with Russian ($z = –71.17$, $p < 0.011$;), Czech ($z = –31.61$, $p < 0.01$) and Slovenian ($z = –20.14$, $p < 0.01$) exhibiting the highest degrees of word order flexibility. In this respect, Estonian follows closely behind the Slavic languages. As noted by Ehala (2006), Estonian is typically classified as an SVO language, but in main clauses, SVX and XVS word orders occur with roughly equal frequency. This flexibility in root contexts is confirmed in our study ($z = –52.82$, $p < 0.01$).

Spoken language corpora reflect increased flexibility in Romance: in this sense similar results can be detected in Spanish ($z = –6.66$, $p < 0.01$) and the two treebanks from French (Rhapsodie 0.11, z = –16.88; ParisStories 0.18, z = –12.99, p < 0.01). The results display an asymmetry with written texts and are in line with the quantitative results of Samo (2025) testing the proposals of Wolfe (2022) for contemporary French, in which the topics are described as base generated directly in the CP.

Encyclopedic entries show less flexibility than heterogeneous corpora in Chinese (z = –41.89), Russian (z = –25.65, p < 0.01), and Dutch (z = –49.20, p < 0.01). In the case of Hebrew (z = –28.09, p < 0.01), their behavior aligns more closely with newswire texts. Knowledge transfer in encyclopedic entries can therefore be seen as a domain where complex configurations are, where possible, avoided in favor of canonical word orders across languages. Similarly, the Romanian medical dataset SiMoNERo supports the findings of Zhao et al. (2021), confirming the crosslinguistic tendency to minimize complexity in medical content (z = –27.38, p < 0.01). On the other hand, other Legal texts crosslinguistically show relatively high flexibility. Finally, the PUD treebanks display similar behavior across languages: given that these are composed of manually curated, often well-formed sentences, this result is relatively expected.

Figure 1 also shows clear distinctions across genres: news vs. heterogeneous sources in Arabic; high flexibility in Chinese legal texts; a well-distributed balance between heterogeneous sources, domain-specific, and PUD data in Czech; clear differences between encyclopedic and news texts in Dutch; and the previously mentioned “Spoken” dimension in Romance languages.

Despite the lack of granularity, this study offers two key contributions. First of all, some trends can be detected. The first of these trends is of pure grammatical nature: Slavic languages show higher flexibility with XP objects with respect to other SVO languages. Estonian behaves similarly. The V2 data from Dutch on the other hand do not show a great variability, and in line with Romance languages (e.g. Italian) that may lead to a criterial approach to the cartography of V2 in non-subject contexts (Samo 2019a).

Secondly, this study introduces a cartographic perspective on text genres and registers based on corpus data, emphasizing how genre may or may not affect (complex) syntactic structures and proposing a quantitative methodology for comparison. Despite limitations in experimentally testing a single phenomenon, the observed trends still offer meaningful syntactic insights. For example, the reduction in syntactic complexity under communicative or clarity pressure supports the broader hypothesis that genre-specific constraints influence the realization of syntactic elements (e.g. encyclopedic entries, vs. spoken).

While the current approach relies on annotated grammatical clauses in treebanks, future work could incorporate sources (cf. Samo & Chen 2022) to investigate such a broad research question. However, as noted by Massaro & Samo (2023) these reordering might be “invisible” in the training of language models, but they can be detected with the right prompting set-up.

We believe that this type of work opens a path for syntacticians to integrate theoretical insights with data-driven methods, further refining syntactic models through simulation and comparison.

5. CONCLUSIONS

In this study, we have refined and extended a methodology initially proposed by Samo et al. (2020) for detecting flexibility of reorderings in syntactically annotated corpora. We adopted a theory-driven, frequency-based approach of genres through corpus analysis, with special attention to morphosyntactically annotated treebanks. We retrieved data from 53 treebanks in 15 languages, across language families.

Further work should extend the scope to a wider range of languages, syntactic configurations, and sources, while ensuring transparency and replicability in automated methods. We argue that quantitative and straightforward computational techniques offer valuable tools for theoretical syntacticians to understand extra-syntactic factors and thus refine syntactic models.

REFERENCES

Aboh, E. O. (2004) *The Morphosyntax of Complement-Head Sequences: Clause Structure and Word Order Patterns in Kwa*, Oxford University Press, Oxford.

Ahrenberg, L. (2007) "LinES: An English-Swedish Parallel Treebank", in J. Nivre, H.-J. Kaalep, K. Muischnek & M. Koit (eds.) *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, Tartu, Estonia, 270–273.

Ahrenberg, L. (2015) "Converting an English–Swedish Parallel Treebank to Universal Dependencies", in J. Nivre & E. Hajičová (eds.) *Proceedings of the Third International Conference on Dependency Linguistics (DepLing 2015)*, Uppsala, Sweden, 10–19.

Alfieri L. & F. Tamburini (2016) "(Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format", in A. Corazza, S. Montemagni & G. Semerano (eds.) *Proceedings of the Third Italian Conference on Computational Linguistics - CLiC-IT 2016*, Accademia University Press, Torino, 19–23.

Androutsopoulos, J. (ed.) (2014) *Mediatization and Sociolinguistic Change*, Mouton de Gruyter, Berlin, Boston.

Behzad, S. & A. Zeldes (2020) "A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging", in A. Barbaresi, F. Bildhauer, R. Schäfer & E. Stemle (eds.) *Proceedings of the 12th Web as Corpus Workshop (WAC-XII)*, European Language Resources Association (ELRA), Paris, 50–56.

Bell, A. (1991) *The Language of News Media*, Blackwell, Oxford, Cambridge, Mass.

Benincà, P. (1995) "Complement Clitics in Medieval Romance. The Tobler-Mussafia Law", in A. Battaye & I. Roberts (eds.) *Clause Structure and Language Change*, Oxford University Press, Oxford, 325–344.

Benincà, P. & C. Poletto (2004) "Topic, Focus, and V2: Defining the CP Sublayers", in L. Rizzi (ed.) *The Structure of CP and IP*, Oxford University Press, Oxford, New York, 52–75.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999) *Longman Grammar of Spoken and Written English*, Pearson Education, Harlow.

Bocci, G. (2013) *The Syntax-Prosody Interface from a Cartographic Perspective: Evidence from Italian,* John Benjamins, Amsterdam, Philadelphia.

Bonilla, J. E. (2024) *Universal Dependencies for Spoken Spanish*, Ghent University, Ghent, Belgium.

Bosco, C., S. Montemagni, & M. Simi (2013) "Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank", in A. Pareja-Lora, M. Liakata & S. Dipper (eds.) *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Association for Computational Linguistics, Sofia, Bulgaria, 61–69.

Bouma, G. & G. van Noord (2017) "Increasing Return on Annotation Investment: The Automatic Construction of a Universal Dependency Treebank for Dutch", in M.-C. de Marneffe, J. Nivre & S. Schuster (eds.) *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, Association for Computational Linguistics, Gothenburg, Sweden, 19–26.

Bybee, J. & S. A. Thompson (2021) "Interaction and Grammar: Predicative Adjective Constructions in English Conversation", *Languages* 7(1), 2, 1–17.

Candito, M., G. Perrier, B. Guillaume, C. Ribeyre, K. Fort, D. Seddah & E. de la Clergerie (2014) "Deep Syntax Annotation of the Sequoia French Treebank", in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 2298–2305.

Cardinaletti, A. (2004) "Toward a Cartography of Subject Positions", in L. Rizzi (ed.) *The Structure of CP and IP, The Cartography of Syntactic Structures, Volume 2*, Oxford University Press, Oxford, 115–165.

Cassidy, L., T. Lynn, J. Barry & J. Foster (2022) "TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish", in S. Muresan, P. Nakov & A. Villavicencio (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 6869–6884.

Cignarella, A. T., C. Bosco, V. Patti & M. Lai (2018) "Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ", in N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan, 7-12 May 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 4204–4211.

Cinque, G. & L. Rizzi (2010) "The Cartography of Syntactic Structures", in B. Heine & H. Narrog (eds.) *Oxford Handbook of Linguistic Analysis*, Oxford, New York, 51–65.

Daniel, T. A. & A. L. Camp (2020) "Emojis Affect Processing Fluency on Social Media", *Psychology of Popular Media* 9(2), 208.

de Marneffe, M. C., C. Manning, J. Nivre & D. Zeman (2021) "Universal Dependencies", *Computational Linguistics* 47(2), 255–308.

di Felippo, A., N. Roman Trevisan, T.A.S. Pardo & L. Panta de Moura (2024) "The DANTEstocks Corpus: An Analysis of the Distribution of Universal Dependencies-Based Part of Speech Tags", *Revista da Abralin* 22(2), 249–271.

Durrleman, S., T. Marinis & J. Franck (2016) "Syntactic Complexity in the Comprehension of Wh-Questions and Relative Clauses in Typical Language Development and Autism", *Applied Psycholinguistics* 37, 1501–1527.

Ehala, M. (2006) "The Word Order of Estonian: Implications", *Journal of Universal Language* 7, 49–89.

Endo, Y. (2022) "Non-Standard Questions in English, German, and Japanese", *Linguistics Vanguard* 8(2), 251–260.

Fontana, J. M. (1997) *Phrase Structure and the Syntax of Clitics in the History of Spanish*, Mouton de Gruyter, Berlin.

Friedmann, N., A. Belletti & L. Rizzi (2009) "Relativised Relatives: Types of Intervention in the Acquisition of A-bar Dependencies", *Lingua* 119, 67–88.

Gamhewage, G., H. Utunen, M. Attias & R. George (2020) "Fast-tracking WHO's COVID-19 Technical Guidance to Training for the Frontline", *Weekly Epidemiological Record* 95, 257–264.

Guillaume, B., M.-C. de Marneffe & G. Perrier (2019) "Conversion et améliorations de corpus du français annotés en Universal Dependencies", *Traitement Automatique des Langues* 60(2), 71–95.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen & M. Baroni (2018) "Colorless Green Recurrent Networks Dream Hierarchically", in M. Walker, H. Ji & A. Stent (eds.) *Proceedings of the 2018 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* Association for Computational Linguistics, New Orleans, Louisiana, 1195–1205.

Haegeman, L. (1997) "Register Variation, Truncation and Subject Omission in English and in French", *English Language and Linguistics* 1, 233–270.

Haegeman, L. (2013) "The Syntax of Registers: Diary Subject Omission and the Privilege of the Root", *Lingua* 130, 88–110.

Haider, H. & L. Szucsich (2022) "Slavic languages – 'SVO' languages without SVO Qualities?", *Theoretical Linguistics* 48(1-2), 1–39.

Hladká, B. V., J. Hajič, J. Hana, J. Hlaváčová, J. Mírovský & J. Votrubec (2007) *Průvodce Českým akademickým korpusem 1.0/ The Czech Academic Corpus 1.0 Guide*, Karolinum, Praha, Czech Republic.

Jelínek, T. (2017) "FicTree: A Manually Annotated Treebank of Czech Fiction", in J. Hlaváčová (ed.) *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017)*, Charles University, Praha, Czech Republic, 181–185.

Kahane, S., B. Caron, E. Strickland & K. Gerdes (2021) "Annotation Guidelines of UD and SUD Treebanks for Spoken Corpora: A Proposal", in D. Dakota, K. Evang & S. Kübler (eds.) *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, Association for Computational Linguistics, Sofia, Bulgaria, 35–47.

Kríž, V. & B. Hladkán (2018) "Czech Legal Text Treebank 2.0", in N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 4501–4505.

Ledgeway, A. (2010) "Subject Licensing in CP: The Neapolitan Double-Subject Construction", in P. Benincà & N. Munaro (eds.) *Mapping the Left Periphery*, Oxford University Press, Oxford, 257–296.

Leung, H., R. Poiret, T. Wong, X. Chen, K. Gerdes & J. Lee (2016) "Developing Universal Dependencies for Mandarin Chinese", in K. Hasida, K.-F. Wong, N. Calzorari & K.-S. Choi (eds.) *Proceedings of the 12th Workshop on Asian Language Resources*, The COLING 2016 Organizing Committee, Osaka, Japan, 20–29.

Margaretha, E. & H. Lüngen (2014) "Building Linguistic Corpora from Wikipedia Articles and Discussions", *Journal for Language Technology and Computational Linguistics* 29(2), 59–82.

Martínez, E., F. Mollica & E. Gibson (2023) "Even Lawyers Do Not Like Legalese", *PNAS* 120(23), e2302672120.

Massaro, A. P. & G. Samo (2023) "Prompting Metalinguistic Awareness in Large Language Models: ChatGPT and Bias Effects on the Grammar of Italian and Italian Varieties", *Verbum* 14, 1–11.

McDonald, R. T., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. B. Hall, S. Petrov, H. Zhang, O. Täckström (2013) "Universal Dependency Annotation for Multilingual Parsing", in H. Schuetze, P. Fung & M. Poesio (eds.) *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Association for Computational Linguistics, Sofia, Bulgaria, 92–97.

Merlo, P. (2016) "Quantitative Computational Syntax: Some Initial Results", *Italian Journal of Computational Linguistics* 2(1), 11–29.

Mititelu, V. B. (2018) “Modern Syntactic Analysis of Romanian”, in O. Ichim, L. Botoşineanu, D. Butnaru, M.-R. Clim, O. Ichim, V. Olariu (eds.) *Clasic și modern în cercetarea filologică românească actuală,* Alexandru Ioan Cuza University Press, Iași, 67–78.

Mititelu, V. B. & M. Mitrofan (2020) “The Romanian Medical Treebank – SiMoNERo”, in G. Singh Lehal, D. Misra Sharma & R. Sangal (eds.) *Proceedings of the 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing,* NLP Association of India, Hyderabad, India, 7–16.

Muischnek, K., K. Müürisep, T. Puolakainen, E. Aedmaa, S. Kirt & D. Särg (2014) “Estonian Dependency Treebank and Its Annotation Scheme”, in V. Henrich, E. Hinrichs, P. Osenove & A. Przepiórkowski (eds.) *Proceedings of the 13th Workshop on Treebanks and Linguistic Theories (TLT13),* University of Tübingen, Germany, 285–291.

Muischnek, K., K. Müürisep & D. Särg (2019) “CG Roots of UD Treebank of Estonian Web Language”, in E. Bick & T. Trosterud (eds.) *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar – Methods, Tools and Applications*, Linköping University Electronic Press, Linköping, Sweden, 23–26.

Nivre, J. (2015) “Towards a Universal Grammar for Natural Language Processing”, in A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing: CICLing 2015, Lecture Notes in Computer Science,* Springer, Cham, 3–16.

Poudat, C., H. Lüngen & L. Herzberg (2024) *Investigating Wikipedia: Linguistic Corpus Building, Exploration and Analysis*, John Benjamins, Amsterdam.

Puskás, G. (2000) *Word Order in Hungarian: The Syntax of A'-positions*, John Benjamins, Amsterdam.

Rademaker, A., F. Chalub, L. Real, C. Freitas, E. Bick & V. de Paiva (2017) “Universal Dependencies for Portuguese”, in S. Montemagni & J. Nivre (eds.) *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, Linköping University Electronic Press, Linköping, Sweden, 197–206.

Rizzi, L. (1982) *Issues in Italian Syntax*, Foris, Dordrecht.

Rizzi, L. (1997) “The Fine Structure of the Left Periphery”, in L. Haegeman (ed.) *Elements of Grammar,* Kluwer, Dordrecht, 281–337.

Rizzi, L. (2004) “Locality and Left Periphery”, in A. Belletti (ed.) *Structures and Beyond: The Cartography of Syntactic Structures.* Oxford University Press, Oxford, New York, 223–251.

Rizzi, L. (2014) “Syntactic Cartography and the Syntacticisation of Scope-Discourse Semantics”, in A. Reboul (ed.) *Mind, Values, and Metaphysics: Philosophical Essays in Honor of Kevin Mulligan* (Vol. 2), Springer, Cham, 517–533.

Rizzi, L. (2016) “Monkey Morpho-Syntax and Merge-Based Systems”, *Theoretical Linguistics* 42(1–2), 139–145.

Rizzi, L. & G. Cinque (2016) “Functional Categories and Syntactic Theory”, *Annual Review of Linguistics* 2, 139–163.

Rüdiger, S. & C. Lange (2023) “Introduction to the Special Issue on ‘The Language of Science Fiction’”, *Linguistics Vanguard* 9(3), 229–232.

Samo, G. (2019a) *A Criterial Approach to the Cartography of V2*, Linguistik Aktuell 257, John Benjamins, Amsterdam, Philadelphia.

Samo, G. (2019b) “Cartography and lLocality in German”, *Rivista di Grammatica Generativa* 63–91.

Samo, G. (2022) “Move to ModP or Base-Generated in FrameP? A Quantitative Cartographic Study”, *Revue Roumaine de Linguistique* LXVII (4), 345–361.

Samo, G. (2023) “Testing Cartographic Proposals on Locality Effects in V2: A Quantitative Study”, *Journal of Historical Syntax* 7(24), 1–32.

Samo, G. (2025) “A Computational Cartographic Study on the Merge Nature of Topics in Stages of French”, *Canadian Journal of Linguistics / Revue canadienne de linguistique*.

Samo, G. & X. Chen (2022) “Syntactic Locality in Chinese In-situ and Ex-situ Wh-Questions in Transformer-Based Deep Neural Network Language Models”, in *Workshop on Computational Linguistics on East Asian Languages (29th International Conference on Head-Driven Phrase Structure Grammar)*, Online.

Samo, G. & E. Isolani (2024) “Comparing Models on the Optionality of Complementizer Omission: A Quantitative Computational Study on German and Italo-Romance”, *Annali di Ca’ Foscari, Serie occidentale* 58, 287–308.

Samo, G. & P. Merlo (2019) “Intervention effects in Object Relatives in English and Italian: A Study in Quantitative Computational Syntax”, in X. Chen & R. Ferrer-i-Cancho (eds.) *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019),* Association for Computational Linguistics, Paris, France, 46–56.

Samo, G. & F. Si (2022) “Optionality of 的 De in Chinese Possessive Structures: A Quantitative Study”, *Quaderni Di Linguistica E Studi Orientali 2022*, 37–53.

Samo, G., Y. Zhao & G. Gamhewage (2020) “Syntactic complexity of Learning Content in Italian for COVID-19 Frontline Responders: A study on WHO’s Emergency Learning Platform”, *Verbum* 11, 4.

Sanguinetti, M. & C. Bosco (2014) “PartTUT: The Turin University Parallel Treebank”, in R. Basili, C. Bosco, R. Delmonte, A. Moschitti & M. Simi (eds.) *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, Springer, Cham, 51–69.

Sanguinetti, M., C. Bosco, A. Lavelli, A. Mazzei, O. Antonelli & F. Tamburini (2018) “PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies”, in N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan, 7-12 May 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 1768–1775.

Si, F. (2011) “Syntacticization of Pragmatic Information at Sentential and Noun Phrase Levels”, *Journal of Yili Normal University (Social Science Edition)* 2, 92–95.

Silveira, N., T. Dozat, M.-C. de Marneffe, S. R. Bowman, M. Connor, J. Bauer & C. D. Manning (2014) “A Gold Standard Dependency Corpus for English”, in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 2897–2904.

Smrž, O., V. Bielický, I. Kouřilová, J. Kráčmar, J. Hajič & P. Zemánek (2008) “Prague Arabic Dependency Treebank: A Word on the Million Words”, in *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008),* European Language Resources Association (ELRA), Marrakech, Morocco, 16–23.

Taulé, M., M. A. Martí & M. Recasens (2008) “AnCora: Multilevel Annotated Corpora for Catalan and Spanish”, in *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008),* European Language Resources Association (ELRA), Marrakech, Morocco, 96–101.

Ursini, F.-A. & G. Samo (2025) “Extracting Toponyms from OpenStreetMap and Other Gazetteers: Comparing Representational Accuracy in Multilingual Contexts”, *Nature – Humanities and Social Sciences Communications*.

van Noord, G., G. Bouma, F. van Eynde, D. de Kok, J. van der Linde, I. Schuurman, E. Tjong Kim Sang & V. Vandeghinste (2013) “Large-Scale Syntactic Annotation of Written

Dutch: Lassy", in P. Spyns & J. Odijk (eds.) *Essential Speech and Language Technology for Dutch*. Springer, Cham, 147–164.

Volk, M., J. Graën & E. Callegaro (2014) "Innovations in Parallel Corpus Search Tools", N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 3172–3178.

Wolfe, S. (2015) *Microvariation in Medieval Romance Syntax: A Comparative Study*, PhD thesis, University of Cambridge.

Wolfe, S. (2022) "Microvariation and Change in the Romance Left Periphery", *Probus* 34(1), 235–272.

Zeldes, A. (2017) "The GUM Corpus: Creating Multilayer Resources in the Classroom", *Language Resources and Evaluation* 51(3), 581–612.

Zeldes, A., N. Howell, N. Ordan & Y. B. Moshe (2022) "A Second Wave of UD Hebrew Treebanking and Cross-Domain Parsing", in Y. Goldberg, Z. Kozareva & Y. Zhang (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates , 4331–4344.

Zhao, Y., G. Samo, H. Utunen, O. Stucke & G. Gamhewage (2021) "Evaluating Complexity of Digital Learning in a Multilingual Context: A Cross-Linguistic Study on WHO's Emergency Learning Platform", *Studies in Health Technology and Informatics,* 516–517.