MOVED TO MODP OR BASE-GENERATED IN FRAMEP? A QUANTITATIVE CARTOGRAPHIC STUDY IN ROMANCE

GIUSEPPE SAMO[1]

Abstract Non-arguments and adverbs are plausible candidates for filling the initial slot of sentences cross-linguistically. Following standard assumptions on the computational operations in syntactic cartography, two possible models can account for adverbs and non-arguments in initial positions, namely a (i) base-generation in a dedicated left peripheral position and (ii) a movement theory. In this paper, we aim to test the generalisation ability of these two models in grammatical clauses exploring quantitative and computational methods. After having discussed a methodology for creating expected counts of grammatical clauses, we test the two models against data extracted from twelve morpho-syntactically annotated treebanks of five Romance languages (French, Italian, Portuguese, Romanian, Spanish). Our results suggest that the (ii) movement theory better captures the data in all languages under investigation. This paper enriches the study exploring the tools of Quantitative Computational Syntax, which aims to test fine-grained theoretical linguistic proposals by exploring large-scale databases.

Keywords: cartography, base-generation, movement theory, Romance languages.

1. INTRODUCTION

The Left Periphery of the clause (henceforth LP; Rizzi 1997) has been under the focus of cartographers (Benincà and Poletto, 2004; Frascarelli and Hinterholzl 2007; Ledgeway 2010; Bianchi et al. 2015, *inter alia*). Adverbs and non-arguments are plausible candidates filling the very beginning of the clause cross-linguistically (cf. Biloa 2013). Let us take, as an initial reference, two sentences from Italian, given in (1), involving the temporal element *domani* 'tomorrow' (1a) and the locative element *in Toscana* 'in Tuscany' (1b).

(1)	a.	*Domani*	*la pittrice*	*ritirerà*	*il*	*premio.*
		tomorrow	the-painter.SG.F	collect.FUT	the	prize
		'Tomorrow, the painter will receive the award'				

[1] Beijing Language and Culture University, samo@blcu.edu.cn.

All the queries and the relevant sentences are available at the following link: https://github.com/samo-g/ModPRomance.

b. *In Toscana* *la pittrice* *ritirerà* *il premio*
in Tuscany the-painter.SG.F collect.FUT the prize
'In Tuscany, the painter will receive the award'

Assuming basic operations of computation of grammatical elements (see Rizzi 2017 and reference therein for a detailed discussion in cartographic terms), there are two logically possible syntactic derivations for the sentences in (1). The temporal and the locative elements in (1a) can be either (i) generated/externally merged in a dedicated position anchoring the spatial-temporal/contextual coordinates of the sentence or (ii) moved/internally merged from the IP targeting dedicated functional projections in the LP.

Both dimensions have been explored within the literature. The model in (i) postulates dedicated functional projections higher than/within the LP, such as FrameP (Benincà and Poletto 2004 for an early discussion; Haegeman and Greco 2016; Wolfe 2019; De Clerq and Haegeman 2021), considered as the locus of generation for initial non-arguments (henceforth INA, singular/plural form). Parallelly, the model in (ii) describes the INA internally merged elements (created within the Inflectional Phrase IP) moved to dedicated criterial positions within the LP (FocusP, TopicP, ModP; Rizzi 2004, Samo 2022)[2]. The two models are summarized in (2).

(2) a. i. [$_{\text{FrameP}}$ *Domani* [$_{\text{CP/IP}}$ *la pittrice ritirerà il premio*]]
ii. [$_{\text{TopicP/FocusP/ModP}}$ *Domani* [$_{\text{IP}}$ *la pittrice <domani> ritirerà il premio*]]
b. i. [$_{\text{FrameP}}$ *In Toscana* [$_{\text{CP/IP}}$ *la pittrice ritirerà il premio*]]
ii. [$_{\text{TopicP/FocusP/ModP}}$ *In Toscana* [$_{\text{IP}}$ *la pittrice <in Toscana> ritirerà il premio*]]

In this paper, we shall present the results of a study testing the generalisation ability of the two cartographic models for grammatical clauses by implementing methods in quantitative computational syntax (in the spirit of Merlo 2015; Merlo and Ouwayda 2018; Gulordava and Merlo 2020, Samo and Merlo 2019, 2021; Merlo and Samo 2022). Following Merlo (2016) and adopting frequency of grammatical clauses as a dependent variable to test linguistic proposals. We first encode the two syntactic proposals as a single feature-based representation and build expected distributions of configurations of grammatical clauses (extracted from corpora) in a predictive regime (in the spirit of Samo and Merlo 2019, 2021).

We test the models drawing from quantitative data, in terms of grammatical clauses. We focus on Romance languages which have been central in the early detection of cartographic functional projections (Rizzi 1997; Cinque 1999; Bonan, 2021 *inter alia* and Rizzi and Samo, 2022 for an overview on the role of Romance languages in cartographic studies). We will operate our counts on twelve treebanks for five languages: French, Italian, Romanian, Spanish, Portuguese.

To reach our goal, we proceed as follows. Section 2 presents the theories under discussion and their predictions on grammatical clauses. In section 3, a quantification of the

[2] In this paper we will not discuss in detail all the other operations involved (e.g., verb movement) with non-argument fronting, such as locative fronting (see Sluckin et al., 2021 for an overview).

models and their theoretical expectations are discussed. Section 4 will present materials and methods, while section 5 discusses the results and some notes on a quantitative (computational) cartographic syntax. Section 6 concludes.

2. MODELLING CARTOGRAPHIC THEORIES

In this section, we present two cartographic models accounting for the syntactic behaviour of fronted adverbials and obliques. We first discuss a model stipulating that INA are the result of a (i) base-generation in dedicated functional projections (Benincà and Poletto 2004; Wolfe 2015, 2019; Haegeman and Greco 2016; De Clerq and Haegeman 2021) followed by the presentation of the model proposing that INA undergo a (ii) criterial movement to the LP (Rizzi 2004; Samo 2019a, 2022) and its predictions in terms of locality (Rizzi 1990, 2004; Starke 2001).

Base-generation of scene setters: Benincà and Poletto (2004) enriched the landscape on studies on Rizzi (1997)'s LP. Benincà and Poletto (2004) suggested that the LP should be rethought in terms of layers. Among the postulated layers, Benincà and Poletto (2004) discussed the existence of a dedicated space in the very top area of the LP labelled as FRAME, in main clauses, locus of generation of syntactic elements which act as "setting the scene" (Benincà and Poletto 2004: 66). The distribution of INA (e.g., temporal elements, but also locative elements) is however restricted to a series of constraints. For example, as given in (3), they possibly fill a position lower than the one dedicated to Hanging Topics (see also Stark, forthcoming and reference therein).

(3) a. *Mario, nel 1999, gli hanno dato il premio Nobel*
Mario in-the 1999 to-him have given the prize Nobel
b. *??Nel 1999, Mario, gli hanno dato il premio Nobel*
in-the 1999 Mario to-him have given the prize Nobel
(Benincà and Poletto 2004; 67, ex. 46a,b)
c. *Mario, a Stoccolma, gli hanno dato il premio Nobel*
Mario in Stockholm to-him have given the prize Nobel
d. *??A Stoccolma, Mario, gli hanno dato il premio Nobel*
in Stockholm Mario to-him have given the prize Nobel

In later studies, the label FrameP has been adopted for the locus of generation higher than the Left Periphery (cf. Wiltschko 2014 *inter alia*). Assuming Haegeman and Greco (2016), in the spirit of Lewis (1975), Frame Setters provide temporal and/or modal restrictions to the set of circumstances of evaluation for the proposition expressed in the main clause. Similarly, Freywald et al. (2013: 12) consider that the leftmost element "fulfills the function of providing an interpretational frame or anchor for the following statement, first, in terms of time, place, condition (in the case of adverbials meaning 'from now on', 'yesterday', 'every year', 'if you are in school' and so on), or second, more abstractly, in terms of discourse linking (as is the case in certain uses of the equivalents of 'then' and 'afterwards')." In syntactic (cartographic) typology, FrameP has been explored in mapping and accounting violations to verb second (V2, cf. Holmberg 2015) in V2

languages (see recent discussions, from two different point of views, in Wolfe 2019 and Samo 2022).

In a nutshell, FrameP is believed to host all those items creating a frame, especially from a temporal, but also spatial point of view. Discussed as a phenomenon strictly delimited to root clauses (Benincà and Poletto 2004: 66), these elements are present given that a syntactic/semantic connection exists. From a modeling point of view, we should not expect any asymmetries related to IP internal configurations (see section 3) since the element is generated higher and no movement, and therefore no locality effects, are involved.

However, the postulation of FrameP creates “unexpected” asymmetries. For example, the postulation of FrameP does not solve “unexpected” crosslinguistically microvariation (see details in Samo 2019a: 85–108; 135–160 for a bigger set of V2 languages) and, intralinguistical optionality, when, for example, the very same "lexical” material is sentence-initial (e.g., *uossa* ‘then’ in both Putèr and Vallader, two Swiss Romansh varieties discussed in details Samo 2022: 158–163). Such asymmetries led Samo (2022) to adopt a movement approach (in the spirit of Rizzi 2004) for INA.

Movement to the LP: INA are plausible candidates to target Topics and Foci positions (Rizzi 1997; Frascarelli and Hinterholzl 2007; Bianchi et al. 2015), if they bear the relevant features (e.g., +Top, +Foc). An additional functional projection, ModP (Rizzi 2004), represents a landing site for modifiers (adverbials) within the LP. We do find neutral contexts in which fronted adverbial are neither topicalized nor focussed (Rizzi 2004; Rizzi and Bocci 2017), but simply “highlighted”, adopting Rizzi’s (2004: 203) terminology. Contrary to topics, highlighted adverbials do not require any connection to a previous background, and they can be uttered as answers to “what happened questions” (Rizzi 2004: 238).

Specifically, this model assumes that INA are generated in the IP domain of the syntactic architecture. Adverbs are plausibly generated in Cinque (1999)’s hierarchy given in (4a), whereas complements/obliques enter the syntactic tree in Schweikert’s (2005) hierarchy (4b).

(4) a. Cinque (1999)’s hierarchy (Cinque 1999: 156)
[MoodSpeech-act [MoodEvaluative [MoodEvidential [ModEpistemic [TPast [TFuture [ModIrrealis [ModNecessity [ModPossibility [AspHabitual [AspRepetetive [AspFrequentative(I) [ModVolitional [AspCelerative(I) [TAnterior [AspTerminative [AspContinuative [AspPerfect(?) [AspRetrospective [AspProximative [AspDurative [AspGeneric/progressive [AspProspective [AspSg.Completive(I) [AspPl.Completive [Voice [AspCelerative(II) [AspRepetetive(II) [AspFrequentative(II) [Aspsg.Completive(II)]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]
b. Schweikert (2005)’s hierarchy (Schweikert 2005:132)
TEMPORAL > LOCATIVE > COMITATIVE > BENEFACTIVE > REASON > SOURCE > GOAL > MALEFACTIVE > INSTRUMENTAL > MATTER > MANNER

Movement implies locality. “Cartographically” speaking, moved elements can be subject to intervention effects (Rizzi 1990; 2004, 2013; Starke 2001). A local relation, in this case a movement to the LP can be disrupted if a potential intervener (an element bearing certain shared properties with the moved object) hierarchically lies between the locus of generation and the landing site. The results of extensive experimental studies show that the dissimilarity in values between the moved element and the intervener helps (or

hinders) adult grammars in parsing grammatical sentences and improves (or disrupts) comprehension in specific populations of speakers, such as child grammar, atypical development or in language pathology (Friedmann et al., 2009; Durrleman et al., 2016; Villata et al., 2016; Chesi and Canal 2019; Martini et al. 2020 *inter alia*).

Hallmarks of movement with fronted non-arguments can be detected. For example, we present in (5a-b) the case when a lower adverb in Cinque's (1999) hierarchy, the celerative adverb *rapidamente* 'rapidly', crosses another adverb which is located higher in the structure (and uttered in the same sentence), such as the higher epistemic adverb *probabilmente* 'probably'. If the lower adverb is "highlighted," the sentence results ungrammatical (5a), whereas if the lower adverb is focussed (thus, dissimilarity in features), the sentence is grammatical (5b). In the case that the higher adverb is "highlighted", no intervention is at play and therefore the movement to ModP results grammatical, as in (5c).

(5) a. *Rapidamente, Gianni ha probabilmente trovato la soluzione.
rapidly Gianni as probably found the solution
solution

b. RAPIDAMENTE, Gianni ha probabilmente trovato
rapidly Gianni has probably found
la soluzione, (non lentamente)
the solution (not slowly)
'Rapidly, Gianni probably found the solution'
(Rizzi and Bocci, 2017, pp. 16–17)

c. Probabilmente, i tecnici hanno risolto rapidamente
probably the technicians have resolved rapidly
il problema
the problem
'Probably, the technicians rapidly resolved the problem.'
(Rizzi 2004, p. 234, 33d)

In a similar vein, if a lower complement crosses a higher complement in Schweikert (2005)'s hierarchy, the sentence is judged better if the lower complement is focussed (e.g., a locative crossing a temporal), as in (6b). Despite being grammatical (6a) "requires some topical properties (on one of the two elements), which we mark with the diacritic #" (Samo 2022: 153–154). Similar to (5c), if the higher complement is the "highlighted" element, the sentence results fully grammatical, as in (6c, to be compared with 1a and 1b where only one item is present).

(6) a. #Alla stazione compravo il giornale alle sei
at.the station buy.IPFV.1s the newspaper at.the six
del mattino
of.the morning

b. ALLA STAZIONE compravo il giornale
at.the station buy.IPFV.1s the newspaper
alle sei del mattino (,non in piazza)
at.the six of.the morning (not in.the square)

'It was at the station that I bought the newspaper at 6am (, not in the square).

c.	Alle sei del mattino	compravo	il giornale
	at.the six of.the morning	buy.IPFV.1s	the newspaper
	alla stazione		
	at.the station		

'I used to buy the newspaper at the station at 6am'

(Samo 2022: 154; 14b,c,a)

The two cartographic proposals can be summarized in Figure 1.

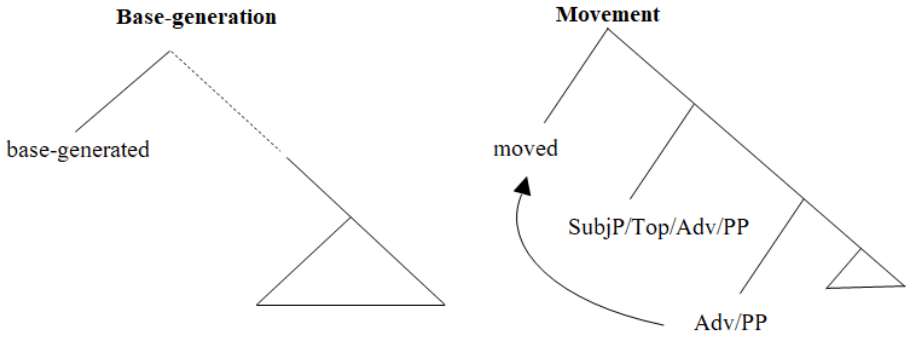

Figure 1.
The two models under investigation

The movement approach predicts some locality issues when the fronted adverbial/PP crosses another element hierarchically higher in the structure.

The base-generation approach does not predict any type of locality effects. On the other hand, the movement model predicts different types of intervention effects, creating ungrammaticality or difficulty in parsing. The fronted adverbial can cross another higher adverb (leading to ungrammaticality) only if other discourse features are involved (e.g., +Focus, +Topics).

The quantitative dimension we want to explore forces us to work on grammatical clauses that we can retrieve from corpora. As Figure 1 (right panel) shows, subject elements (cf. Cardinaletti 2004; Rizzi 2015) moved from the vP area can also be plausible interveners in terms of locality if INA is moved. Subjects are distinct elements from INA, since INA tendentially do not bear any person feature and not every INA is marked with number/gender feature. However, following Samo and Merlo (2021: 23), a feature that could result in intervention effects of the subjects with all INA (adverbials, arguments) in grammatical clauses is a feature labelled as TYPE (XP *vs. pronominal/null*). Subjects can be realized as maximal projections (e.g. *la professoressa* 'the professor'), pronominal heads (e.g., *lei* 'she') or, in certain languages such as Italian, by a null element when certain requirements are satisfied (Rizzi 1982; Frascarelli 2007). Following Cinque (1999) and Schweikert (2005), we consider adverbs and non-clitic obliques as maximal projections (XP) since they fill Spec positions within the syntactic architecture. Moreover, studies in processing and locality showed that mismatch in PP does not work as an ameliorative effect

(cf. Costa et al. 2014). We operate on how different INA and subjects in a given configuration are, and, therefore, the main feature under investigation is the maximal projection/lexical nature +N/-N (following the label in Rizzi 2018: 348). We quantify our hypotheses in grammatical clauses in section 3.

3. QUANTIFYING THE HYPOTHESIS

Our measures are represented by the frequency of grammatical clauses in syntactically annotated treebanks and the distributions of opposite patterns (+N, –N). The interaction between frequency of grammatical structures and grammar has been highly debated in the literature (see Yang et al. 2017 and Ibbotson 2013 for different overviews on frequency and grammar). Large-scale datasets (see Nivre 2015 *inter alia*) however allow "us to develop investigations of the correlation between quantitative linguistic properties and theory-driven abstract linguistic representations and operations." (Samo and Merlo 2021: 29). In other words, we follow Merlo (2016) and related works in Quantitative Computational Syntax by using quantitative measures as a dependent variable to test the linguistic proposals (and therefore grammatical properties) under investigation. We mainly investigate distributions of structures in treebanks cross-linguistically.

As stated in section 2, the base-generation approach does not make clear predictions, while the movement model foresees a series of locality issues. The main element we take into consideration is the nature of the subject: maximal projection ($Subj_{XP}$), pronominal ($Subj_{Pro}$) and null subjects (NS). According to Belletti and Rizzi (2013) pronouns represent to a lesser extent interveners because of their lack of a lexical restriction (in our terms, generalizing, a +N feature). Table 2 summarizes the predictions of the two models in terms of locality between the INA and the nature of the intervening subject.

Table 2.

The two models under investigation and their prediction in terms of locality

STRUCTURE	CONFIGURATION	BASE-GENERATION	MOVEMENT
INA – $Subj_{XP}$	+N	No intervention	Intervention
INA – $Subj_{Pro}$ / NS	-N	No intervention	No Intervention

Based on Table 2, we can postulate the hypotheses. The impact of the intervention (created by movement) needs to be compared with an imputed distributions of the nature of the subject ($Subj_{XP}$, $Subj_{Pro}$, NS) in standard/canonical clauses when no movement is at play[3]. We compare the observed distributions to those predicted by the two models. Our alternative hypothesis is represented by the *movement theory* since it makes clear predictions on the intervention effects between INA and subjects.

[3] The readers is referred to the extended discussion, methodologically and theoretically, in Samo and Merlo (2021: 11–14; 23–25) with respect to the creation of "expected" counts on the basis of distribution of features in canonical clauses.

H1: The distributions of –N/+N configurations of observed counts with INA should be closer to the predicted distributions of the movement theory in which mismatching configurations (–N) are favoured.

Following Merlo and Samo (2022), we measure the distance of distributions of syntactic configurations (+N, –N) via the Kullback-Leibler divergence (D_{KL}). D_{KL} represents a measure of how one probability distribution *P* is different from a second reference probability distribution *Q*, calculated as follows:

$$D_{KL}\left(P \parallel Q\right) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

We investigate, as a probability distribution, the configurations in which the subject bears a +N or a -N feature. In our case, the reference distribution is the observed distribution (*Obs*). Let *Obs* be the observed distribution, *Bg* (Base-generation) be the expected distribution of +/– features in canonical clauses and *Move* (movement) the distribution of features of canonical context with at least minimal advantage to –N configurations (as expected by the theory).

For example, if a fictional distribution for canonical clauses is, e.g., (+N = 0.600; –N = 0.400), these distributions are adopted for the base-generation model (*Bg*), whereas, at least, a marginal advantage is predicted to –N constructions by the theory postulating movement (e.g, + 0.001; then +N = 0.599; –N = 0.401)[4]. Notice that if the two distributions are equal, the D_{KL} is zero. D_{KL} is positive if the distributions are different. The smaller D_{KL}, the closer two distributions are.

Thus, hypothesis H_1 can be formulated as follows:

$$H_1 : D_{KL}(Move \parallel Obs) < D_{KL}(Bg \parallel Obs)$$

Section 4 presents the materials and methods of our quantitative study, while section 5 discusses the results.

[4] It is important to remark that we do not operate on any control group. However, two candidate configurations, which has been well studied (also computationally) would be the case of (i) non-intervention, such as subject relative clauses, and a case of (ii) intervention of arguments, such object crossing subjects in object relative clauses (see Friedmann et al. 2009). However, (i) could not be properly adopted as a control group since there is no null subject relative clauses. Furthermore, the head of a relative clause is tendentially avoided as a head (see Samo and Merlo 2019's 'adjusted counts' models). Another problem is given by the different sizes of subject and object relatives we can found in corpora (e.g., Roland et al. 2007; Belletti and Chesi 2014). A preliminary observation made for this study found that the Romanian treebank SiMoNero (Barbu Mititelu and Mitrofan 2020) which contains medical texts, shows a very reduced number of object relative clauses. This is in line with the results in Zhao et al. (2021)'s work on health-emergency corpora in eight languages (Chinese, English, French, Italian, Polish, Portuguese, Russian, Spanish) in which relatives crosslinguistically are minimized in medical domain content and when they are present there is a preference for subject relatives (see also Samo et al. 2020).

4. MATERIALS AND METHODS

The quantitative evidence is extracted from grammatical clauses extracted from twelve Universal Dependencies treebanks (UD; Nivre 2015, Zeman et al. 2022) in five languages (French, Italian, Spanish, Portuguese, Romanian). Table 2 reports treebanks (and related reference), genres and size of every treebank.

Table 2

Languages, treebanks (version and references), genres and sizes
[A] https://github.com/UniversalDependencies/UD_French-Rhapsodie/ (last access, 09.06.2022)
[B] https://github.com/UniversalDependencies/UD_Portuguese-Bosque (last access, 09.06.2022)
[C] https://github.com/UniversalDependencies/UD_Portuguese-GSD/ (last access, 09.06.2022)

LANGUAGE	TREEBANK	GENRES	TOKENS	SENTENCES
French	GSD v. 2.9 (Guillame et al. 2019)	blog, news, reviews, wiki	389,208	16,341
	Rhapsodie v. 2.9 [A]	spoken	43,700	3,210
	Sequoia (Perrier and al. 2014)	medical, news, non-fiction, wiki	68,596	3,099
Italian	ISDT 2.9 (Bosco et al. 2014)	wiki, legal, news	278,466	14,167
	VIT 2.9 (Alfieri and Tamburini 2016)	news, non-fiction	259,203	10,087
	PoSTWITA 2.9 (Sanguinetti et al. 2018)	social media	119,342	6,712
Romanian	RRT 2.9 (Barbu Mititelu 2018)	academic, fiction, legal, medical, news, nonfiction, wiki	218,510	9,524
	SiMoNero 2.9 (Barbu Mititelu and Mitrofan 2020)	medical	146,020	4,681
Spanish	AnCora 2.9 (Taulé et al. 2008)	news	555,315	17,662
	GSD 2.9 (Bohnet et al. 2013)	blog, news, reviews, wiki	423,344	16,013
Portuguese	Bosque 2.9 [B]	news	210,959	9,357
	GSD 2.9 [C]	blog, news	297,057	12,019

Since different treebanks have different (and multiple) genres and different sizes, we have decided not to provide any comparison and only work at the crosslinguistic dimension (French, Italian, Portuguese, Romanian and Spanish).

All the sentences are extracted with the Grew-match tool maintained by Inria in Nancy (http://universal.grew.fr, last access 05/2022). The queries searched for an element (tendentially punctuation in UD) governing a "root" dependency to the main clause verb. A variable *x* annotated as an adverb in terms of Part-of-Speech (PoS) or as *oblique* dependent to the verb has been required as the first constituent of the main clauses. This variable *x* precedes a variable *y* dependent of a subject dependency, and we manipulated the nature of the element in terms of PoS. Nouns and proper nouns were considered maximal projections ($Subj_{XP}$), while pronominal entities, naturally as pronouns ($Subj_{PRO}$)[5].

A query also detected those contexts in which an adverb and an oblique are initial sentences in main clauses without an overt subject (NS). In a nutshell, the queries retrieved non-arguments (*oblique/advmod* syntactic relations, *adverb/preposition* part-of-speech) as first element of the sentence (INA) and selected nature (maximal projection, pronominal, null) of the subjects. All the queries are available as supplementary files. Some relevant examples are given in Table 3[5].

[5] Cartographically speaking, we need to stress the status of postverbal subjects as interveners in locality when an INA is fronted. Microvariation is at play in Romance postverbal subjects, both crosslinguistically (cf. Martins 2020) and intralinguistically (cf. Cardinaletti 2018). A standard cartographic assumption translates postverbal subjects in Romance as instances of movement to the the LowIP area (cf. Belletti 2004). There is no agreement with respect to the exact locus of the cartographic architecture where Cinque (1999) and Schweikert's (2005) hierarchies merge together (see Rizzi and Cinque 2016: 151) and a fine-grained description in how they interact with the portion of the structure dedicated to Topic/Focus in the IP discussed by Belletti (2004) crosslinguistically. Martins (2020: 109) shows that postverbal pronominal subjects in European Portuguese are rated better if they precede elements like rapidly which represent a lower part of Cinque's (1999) hierarchy. Around this position are plausibly located certain elements of Schweikert's hierarchy (cf. Hinterhölzl 2000). In this respect, postverbal subjects do act as interveners for the movement of the INA to the LP. To confirm this datum, we run a quantitative test for canonical orders discussed in detail in Samo (2019b). More frequent patterns are assigned with the property of "being canonical". We analysed the results extracted from the Italian ISDT v.2.10 treebank, querying occurrences of post-verbal subjects preceding or following obliques, which are not in non-initial position. Also in this case, we explored the interface of Grew-match (http://universal.grew.fr/, last access 11.11.2022). The relevant queries (*pattern {verb-[nsubj]-> subj; verb-[obl]->obl; verb << subj; subj << obl; obl << verb}* and *pattern {verb-[nsubj]-> subj; verb-[obl]->obl; verb << subj; obl << subj; obl << verb}*) retrieved 222 occurrences for postverbal subjects preceding obliques and 151 occurrences of postverbal subjects following obliques. This observed preference is in line with the idea that complements may be generated below the LowIP area. Future studies, in both experimental settings and with quantitative and computational methods, should explore in details the interactions and map the "borders" between portions of the syntactic architecture.

Table 3

Examples of sentences (with ID) of the conditions and queries.

Matching	Configuration	Language	Treebank, ID	Sentences
–N	Adv – NS	Italian	ISDT, 2_Europarl-412	*Adesso siamo già molto in ritardo* 'Now, we have been already much late'
	Obl – NS	Spanish	AnCora, 3LB-CAST-n1-8-s13	*Con eso demostraba su desorientación* 'With that, (she/he/it) demonstrated her/his/its disorientation.'
	Adv – $Subj_{Pro}$	French	GSD, fr-ud-train_02188	*Ensuite ils doivent passer un examen* 'Afterwards, they need to pass an exam'
	Obl – $Subj_{Pro}$	Italian	ISDT, tut-1237	*In questo caso egli risponde nei limiti di quanto ha ricevuto* 'In this case, he responds within the limits of what he has received'
+N	Adv – $Subj_{XP}$	Portuguese	GSD, train-s8555	*Provavelmente a maior performance do grupo foi no Wembley Stadium reformado em 16 e 17 de junho de 2007.* 'Probably the group's biggest performance was at the renovated Wembley Stadium on June 16 and 17, 2007'.
	Obl – $Subj_{XP}$	Romanian	RRT, test-345	*în emisfera nordică această perioadă se întinde din noiembrie până în aprilie.* 'In the northern hemisphere this period runs from November to April'.

We run all the queries in every language. Albeit French is typologically described as a non-null subject language (cf. Dryer 2013), we did find occurrences of null subject elements in spoken corpora such as *mais faut aussi être réaliste* 'but it is necessary to be realist' (Rhapsodie, Rhap_D0006-99).

Although the calculation has been done separately, we will merge the results altogether. The annotation scheme of UD, being extremely compacted to favor crosslinguistic comparison (see the discussion in Samo 2019b about transforming universal dependencies into cartographic representation) does not capture extremely fine-grained distinctions. As a matter of fact, it is possible to find elements that are annotated as adverbials which are obviously complements and vice-versa. For example, the query "Adv – SubjXP" in Spanish can provide target sentences of the type *Ciertamente la demanda de créditos es enorme* 'Certainly, the credit demand is huge' (AnCora, 3LB-CAST-n1-8-s13), but also non-target (e.g., obliques) such as the sentence *Después de*

los fracasos con Lippi y Tardelli, la sensatez que ha traído Cúper supone ahora la gran esperanza para un club que no gana el scudetto desde 1989 'After the failures with Lippi and Tardelli, the good sense that Cúper has brought is now the great hope for a club that has not won the scudetto [Italian football league] since 1989' (AnCora, CESS-CAST-P-20011002-160-s14). INA followed by comma were also taken into consideration, given that 'comma intonation' has also been adopted as referring to internally merged elements like Topics (Rizzi 1997: 285). Naturally, the nature of the data would not provide clear clues concerning the length of a possible phonological break after the INA. The data could benefit of a manual analysis to map more refined asymmetries. The goal of our paper is to keep the automatization process clear for its replicability crosslinguistically, therefore we would like to leave a detailed manual analysis to future studies.

The prior counts on canonical orders are based on the entire distribution available in the relation tables in Grew-Match. We considered XPs all the dependents of a *nsubj* (subject) relation annotated with NOUN (nouns) and PROPN (proper nouns). We considered pronouns all the subjects marked as PRON (pronominal entities). Null subjects were retrieved with a dedicated query which searched all the sentences in the treebank in which a verb inflected in a finite form does not govern any *nsubj* relations. All queries and relation tables are provided in the supplementary files.

Results are presented in section 5.

5. RESULTS AND DISCUSSION

Table 4 summarizes the results with the distribution. Detailed raw numbers, as well all the sentences, are however available in the supplementary files.

Table 4

Languages, total of counts and distributions of conditions.

Language	INA			Canonical		
	Tot	+N	–N	Tot	+N	–N
French	533	0.353	0.647	26997	0.479	0.521
Italian	6098	0.474	0.526	40359	0.499	0.501
Portuguese	1118	0.412	0.588	28493	0.636	0.364
Romanian	1096	0.451	0.549	18503	0.660	0.340
Spanish	4603	0.441	0.559	59552	0.592	0.408

The results in Table *4* visually suggest to us that there is an asymmetry between the distribution of the nature of subjects in every language under investigation. In all languages, the –N conditions are a preferred option. We can turn now to test hypothesis 1 (*H*1) here repeated.

$$H_1 : \mathrm{D_{KL}}(Move \,||\, \mathrm{Obs}) < \mathrm{D_{KL}}(Bg \,||\, \mathrm{Obs})$$

All analyses were conducted using the *dlookr* package (Ryu 2021) in R (R Core Team 2021). Table 5 summarizes the results.

Table 5
Languages and Kullback-Leibler divergence in the two tested environments.
Results in bold confirm the hypothesis H_1.

Language	D_{KL} (*Move* || Obs)	D_{KL} (*Bg* || Obs)
French	**0.144902**	0.146206
Italian	**0.024598**	0.025648
Portuguese	**0.274702**	0.276135
Romanian	**0.249930**	0.251310
Spanish	**0.167701**	0.174321

Our results confirm H_1 for every language under investigation. The model postulating that INA move rather than merge seem to better capture the frequencies of the grammatical clauses. We postulated a minimal advantage, by only giving a +0.001 advantage in terms of distribution, but future studies can quantify the "weight" of operations extracted from the theory (see in this respect, Merlo and Ouwayda 2018 on Universal 20). What we observe is a clear trend which also confirms the discussion in Samo (2019a) for V2 languages and Samo (2022) for Italian and Swiss Romansh, in which INA can plausibly only be considered as a fronted constituent.

Inspired by Samo and Merlo (2019, 2021) and Merlo and Samo (2022), we presented and discussed a simple computational model to predict frequencies of grammatical clauses based on two formal syntactic (in this case, cartographic) proposals. The intra-linguistic variations, in terms of genres/registers, needs to be investigated in treebanks that can be statistically compared. Following Samo and Merlo (2021: 29), "this methodology assumes that underlying grammatical properties surface quantitatively, once independent influences of use are properly factored out." In other words, the predictions made by linguistic proposals represent independent variables and the predicted counts are thus the dependent variable to test the generalization ability of the models.

A final contribution of this work is that we provided a relatively simple and precise methodology that can be used as a "blueprint" for many other theory-driven quantitative studies. As discussed in Samo (2019b), there is no direct mapping with cartographic representations and syntactically annotated corpora. The creation of "cartographically" annotated treebanks will definitely require an excessive workload, while a translation, even only in formal terms like the one presented here, of existing annotation schemata could help in testing, quantitatively and computationally, cartographic proposals.

We believe that this type of methodology could represent a first step as well with respect to the scouting phase in experimental procedures (cf. Goodall (eds) 2021). Naturally, the quantitative dimension does not exclude the important role of judgements, but the statistical results, in our case, could lead to the understanding of the learnability of models and their generalization ability crosslinguistically.

Future studies will refine the methodology and hopefully work also on raw, non-annotated data. In this respect, recent studies investigated locality and featural Relativized Minimality with learning algorithms trained on a large number of non-annotated corpora (e.g. word embeddings; Merlo 2019; Merlo and Ackermann 2018; transformer-based deep neural network language models, Samo and Merlo, to appear). Quantitative and computational analyses represent a tool that, in our opinion, should be harvested by theoretical syntacticians as additional evidence.

6. CONCLUSIONS

In this paper we aimed to develop a quantitative and computational dimension to the established qualitative dimensions of the cartographic representation. Given a specific syntactic configuration, we presented two "rival" theories in accounting the phenomenon. In particular, the two models made different predictions in terms of locality. Exploring the tools of Quantitative Computational Syntax, we tested the generalisation ability of the two models. Our results suggest that the movement for initial non-arguments (INA) approach (Samo 2022) better capture the observed counts crosslinguistically in Romance. Future studies shall refine the methodology and enlarge the number of investigated syntactic configurations and languages.

REFERENCES

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. Semeraro (eds), *Proceedings of the Third Italian Conference on Computational Linguistics - CLiC-IT 2016,* Napoli, 5-6 December 2016, 19–23.

Barbu Mititelu, V., 2018, "Modern Syntactic Analysis of Romanian", in O. Ichim, L. Botoşineanu, D. Butnaru, M.R. Clim, V. Olariu (eds), *Clasic şi modern în cercetarea filologică românească actuală*, Iaşi, Publishing House of "Alexandru Ioan Cuza" University, 67–78.

Barbu Mititelu, V., M. Mitrofan, 2020, "The Romanian Medical Treebank – SiMoNERo", in *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing* – ConsILR-2020, 7–16.

Belletti, A., 2004. "Aspects of the low IP area", in L. Rizzi (ed.), *The Structure of CP and IP [The Cartography of Syntactic Structures, Volume 2*], New York, Oxford University Press, 16–51.

Belletti, A., C. Chesi, 2014, "A syntactic approach toward the interpretation of some distributional frequencies: Comparing relative clauses in Italian corpora and in elicited production", *Rivista di Grammatica Generativa*, 36, 1–28.

Belletti, A., L. Rizzi, 2013, "Intervention in grammar and processing", in I. Caponigro, C. Cecchetto (eds), *From Grammar to Meaning: The Spontaneous Logicality of Language*, Cambridge, Cambridge University Press, 293–311.

Benincà, P., C. Poletto, 2004, "Topic, Focus, and V2: Defining the CP Sublayers", in L. Rizzi (ed.) *The Structure of CP and IP [The Cartography of Syntactic Structures, Volume 2]*, Oxford and New York: Oxford University Press, 52–75.

Bianchi, V., G. Bocci, S. Cruschina, 2015, "Focus fronting and its implicatures", in E.O. Aboh, J. Schaeffer, P. Sleeman (eds) *Romance Languages and Linguistic Theory 2013: Selected papers from 'Going Romance' Amsterdam 2013 [Romance Languages and Linguistic Theory 8],* Amsterdam: John Benjamins, 1–20.

Biloa, E., 2013. *The syntax of Tuki: a cartographic approach*, Amsterdam, John Benjamins Publishing Company.

Bohnet, B., J. Nivre, I. Boguslavsky, R. Farkas, F. Ginter, J. Hajič, 2013, "Joint morphological and syntactic analysis for richly inflected languages", *Transactions of the Association for Computational Linguistics*, 1, 415–428.

Bonan, C., 2021, *Romance Interrogative Syntax: Formal and Typological Dimensions of Variation*. Amsterdam, John Benjamins Publishing Company.

Bosco, C., F. Dell'Orletta, S. Montemagni, M. Sanguinetti, M. Simi., 2014, "The evalita 2014 dependency parsing task", in *EVALITA 2014 evaluation of NLP and speech tools for Italian*, Pisa, Pisa University Press, 1–8.

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 and New York, Oxford University Press, 115–165.

Cardinaletti, A. 2018. "On different types of postverbal subjects in Italian", *Italian Journal of Linguistic* 30, 2, 79–106.

Chesi, C., P. Canal, 2019, "Person features and lexical restrictions in Italian clefts", *Frontiers in Psychology*, 10, 2105.

Cinque, G., 1999, *Adverbs and Functional heads. A cross-linguistic perspective*. New York/Oxford, Oxford University Press.

Costa, J., N. Friedmann, C. Silva, M. Yachini, 2014, "The boy that the chef cooked: Acquisition of PP relatives in European Portuguese and Hebrew", *Lingua*, 150, 386–409.

De Clercq, K. L. Haegeman, 2021, "Invariant die and adverbial resumption in the Ghent dialect", in F. Si, L. Rizzi (eds), *Current issues in syntactic cartography : a crosslinguistic perspective*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 53–110.

Dryer, M.S., 2013, "Expression of Pronominal Subjects", in: M.S. Dryer, M. Haspelmath (eds) *The World Atlas of Language Structures Online*. Leipzig, Max Planck Institute for Evolutionary Anthropology (Available online at http://wals.info/chapter/101, Accessed on 18.06.2022).

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, 6, 1501–1527.

Frascarelli, M., 2007, "Subjects, topics and the interpretation of referential *pro*", *Natural Language and Linguistic Theory*, 25, 4, 691–734.

Frascarelli, M., R. Hinterhölzl, 2007, "Types of Topics in German and Italian", in K. Schwabe, S. Winkler (eds), *On Information Structure, Meaning and Form*, Amsterdam and Philadelphia, John Benjamins, 87–116.

Freywald, U., L. Cornips, N. Ganuza, I. Nistov, T. Opsahl, 2013, "Urban vernaculars in contemporary northern Europe: Innovative variants of V2 in Germany, Norway and Sweden", *Working papers in Urban language and literacies*, 119, 1-21.

Friedmann, N., A. Belletti, L. Rizzi, 2009, "Relativised relatives: Types of intervention in the acquisition of A-bar dependencies", *Lingua*, 119: 67–88.

Goodall, G. (ed.), 2021, *The Cambridge Handbook of Experimental Syntax*, Cambridge, Cambridge University Press.

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, 71–95.

Gulordava, K., P. Merlo, 2020, "Computational quantitative syntax: The case of Universal 18", in *Romance Languages and Linguistic Theory 16: Selected papers from the 47th Linguistic Symposium on Romance Languages (LSRL)*, vol. 16, 109, Newark, Delaware, John Benjamins.

Haegeman, L., C. Greco, 2016, *Frame setters and the microvariation of subject-initial V2*, Ms. University of Ghent, <https://ling.auf.net/lingbuzz/003226> (December 2016).

Hinterhölzl, R., 2000, "Licensing movement and stranding in the West Germanic OV-languages", in P. Svenonius (ed.), *The derivation of VO and OV*, Amsterdam, Benjamins, 293–326.

Holmberg, A., 2015, "Verb Second", in T. Kiss, A. Alexiadou (eds), *Syntax – Theory and Analysis*, Berlin, Walter de Gruyter, 242–283.

Ibbotson, P., 2013, "The Scope of Usage-Based Theory", *Frontiers in Psychology*, 4, 255.

Ledgeway, A., 2010, "Subject Licensing in CP: The Neapolitan Double-Subject Construction", in P. Benincà, N. Munaro (eds) *Mapping the Left Periphery* [The Cartography of Syntactic Structures, Volume 5], Oxford/New York, Oxford University Press, 257–296.

Lewis, D., 1975, "Adverbs of quantification", in E. Keenan (ed.), *Formal semantics of natural language*, Cambridge, Cambridge University Press, 3–15.

Martini, K., A. Belletti, S. Centorrino, M. Garraffa, 2020, "Syntactic complexity in the presence of an intervener: the case of an Italian speaker with anomia", *Aphasiology*, 34, 8, 1016–1042.

Martins, A. M., 2020, “Some notes on Postverbal Subjects In Declarative (And Other Non Wh-) Sentences”, *Revista Diadorim*, 22, 3, 98–119.
Merlo, P., 2015, “Predicting word order universals”, *Journal of Language Modelling*, 3, 2, 317–344.
Merlo, P., 2016, “Quantitative Computational Syntax: Some Initial Results”, *Italian Journal of Computational Linguistics*, 2, 1, 11–30.
Merlo, P., 2019, “Probing word and sentence embeddings for long-distance dependencies effects in French and English”, in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy, Association for Computational Linguistics, 158–172.
Merlo, P., F. Ackermann, 2018, “Vectorial semantic spaces do not encode human judgments of intervention similarity”, in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, Brussels, Belgium, Association for Computational Linguistics, 392–401.
Merlo, P., S. Ouwayda, 2018, “Movement and structure effects on Universal 20 word order frequencies: A quantitative study”, *Glossa: a journal of general linguistics*, 3, 1, 84 doi: https://doi.org/10.5334/gjgl.149.
Merlo, P., G. Samo, 2022, “Exploring T3 languages with quantitative computational syntax”, *Theoretical Linguistics*, 48, 1-2, 73–83.
Nivre, J., 2015, “Towards a Universal Grammar for Natural Language Processing”, in A. Gelbukh (ed.), *International Conference on Intelligent Text Processing and Computational Linguistics*, Cham: Springer, 3–16.
Perrier, G., M. Candito, B. Guillaume, C. Ribeyre, K. Fort, D. Seddah, 2014, “Un schéma d'annotation en dépendances syntaxiques profondes pour le français”, *Proceedings of TALN 2014*, Marseille, France.
R Core Team, 2021, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Rizzi, L., 1982, *Issues in Italian Syntax*, Dordrecht, Foris.
Rizzi, L., 1990, *Relativized Minimality*, Cambridge MA, MIT Press.
Rizzi, L.,1997, “The Fine Structure of the Left Periphery”, in L. Haegeman (ed), *Elements of Grammar*, Dordrecht, Kluwer Academic Publisher, 281–337.
Rizzi, L., 2004, “Locality and Left Periphery”, in A. Belletti (ed.), *Structures and Beyond [The Cartography of Syntactic Structures, Volume 3]*, Oxford, Oxford University Press, 223–251.
Rizzi, L., 2013, “Locality”, *Lingua*, 130, 169–186.
Rizzi, L., 2015, “Notes on Labeling and Subject Positions”, in E. Di Domenico, C. Hamann, S. Matteini (eds) *Structures, Strategies and beyond – Studies in Honour of Adriana Belletti*, Amsterdam, John Benjamins Publishing Company, 17–46.
Rizzi, L., 2017, “On the format and locus of parameters: The role of morphosyntactic features”, *Linguistic Analysis*, 41, 159–191.
Rizzi, L., 2018, “Intervention effects in grammar and language acquisition”, *Probus*, 30, 2, 339–367.
Rizzi, L., G. Bocci, 2017, “The Left Periphery of the Clause: Primarily Illustrated for Italian”, in M. Everaert, H. van Riemsijk (eds), *The Blackwell Companion to Syntax*, Hoboken, John Wiley and Sons, 1–30.
Rizzi, L., G. Samo, 2022, “Introduction: On the Role of Romance in Cartographic Studies”, *Probus*, 1–8.
Roland D., F. Dick, J. L. Elman, 2007, “Frequency of basic English grammatical structures: A corpus analysis”, *Journal of memory and* language, 57, 3, 348–379.
Ryu, C., 2021, *dlookr: Tools for Data Diagnosis, Exploration, Transformation*. R package version 0.5.4.9000, https://CRAN.R-project.org/package=dlookr.
Sanguinetti, M., C. Bosco, A. Lavelli, A. Mazzei, F. Tamburini, 2018, “PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies”, *Proceedings of the LREC 2018*.
Samo, G., 2019a, *A Criterial approach to the Cartography of V2*, Amsterdam, John Benjamins Publishing Company.

Samo, G., 2019b, "Cartography and Locality in German: a quantitative study with Dependency Structures", *Rivista Di Grammatica Generativa/Research in Generative Grammar*, 5, 1–26.

Samo, G. 2022, "Criterial V2: ModP as a locus of microvariation in Swiss Romansh varieties", *Probus*, 1–28.

Samo, G., P. Merlo, 2019, "Intervention effects in object relatives in English and Italian: a study in quantitative computational syntax". In *Proceedings of the first workshop on quantitative syntax* (Quasy, Syntaxfest 2019), Paris, France, Association for Computational Linguistics, 46–56.

Samo, G., P. Merlo, 2021, "Intervention effects in clefts: a study in quantitative computational syntax", *Glossa: a journal of general linguistics*, 6, 1.

Samo, G., P. Merlo, forthcoming, "Distributed computational models of intervention effects: A study on cleft structures in French", in C. Bonan, A. Ledgeway (eds), *It-Clefts: Empirical and Theoretical Surveys and Advances*, De Gruyter.

Samo, G., Zhao, U. Y., & Gamhewage, G., 2020. "Syntactic Complexity of Learning Content in Italian for COVID-19 Frontline Responders: A Study on WHO's Emergency Learning Platform", *Verbum*, 11, 1–4.

Schweikert, W., 2005, *The Order of Prepositional Phrases in the Structure of the Clause*, Amsterdam, John Benjamins Publishing Company.

Sluckin, B. L., S. Cruschina, F.Martin, 2021, "Locative inversion in Germanic and Romance: A conspiracy theory", in S. Wolfe, C. Mecklenborg (eds), *Continuity and Variation in Germanic and Romance*, Oxford, Oxford University Press.

Stark, E., forthcoming, "Hanging topics and frames in the Romance languages: syntax, discourse, diachrony", *Oxford Research Encyclopedia of Linguistics*, Oxford: Oxford University Press.

Starke, M., 2001, *Move dissolves into Merge: a theory of locality*, PhD Dissertation, University of Geneva.

Taulé, M., M.A. Martí, M. Recasens, 2008, "Ancora: Multilevel Annotated Corpora for Catalan and Spanish", *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh (Morocco).

Villata, S., L. Rizzi, J. Franck, 2016, "Intervention effects and relativized minimality: New experimental evidence from graded judgments", *Lingua*, 179, 76–96.

Wiltschko, M., 2014, *The Universal Structure of Categories: Towards a Formal Typology*, Cambridge, Cambridge University Press.

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

Wolfe, S., 2019, "Redefining the typology of V2 languages. The view from Medieval Romance and beyond", *Linguistic Variation*, 19, 16–46.

Yang, C., S. Crain, R.C. Berwick, N. Chomsky, J.J. Bolhuis, 2017, "The growth of language: Universal Grammar, experience, and principles of computation", *Neuroscience and Biobehavioral Reviews*, 81, 103–119.

Zeman, D. *et al.*, 2022, Universal Dependencies 2.9, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1–4611.

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", in J. Mantas *et al.* (eds), *Public Health and Informatics*, vol. CCLXXXI of Studies of Health Technologies and Informatics, Amsterdam, IOS Press, 516–517.