Quaderni di Linguistica e Studi Orientali- QULSO
Working papers in Linguistics and Oriental Studies

Firenze University Press
https://oajournals.fupress.net/index.php/bsfm-qulso

OPEN ACCESS

Citation: G. Samo (2021) N-merge systems in adult and child grammars: a quantitative study on external arguments. *Qulso* 7: pp. 103-130. doi: http://dx.doi.org/10.13128/QUSO-2421-7220-12005

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

N-merge systems in adult and child grammars: a quantitative study on external arguments

Giuseppe Samo
Beijing Language and Culture University (<samo@blcu.edu.cn>)

Abstract:

This paper explores quantitative results based on theoretical assumptions related to the predictions on N-merge systems (Rizzi 2016) ranked from minimum to a maximum of complexity in terms of the computational devices and derivational operations they require. We investigate the nature of external arguments focussing on 2-merge systems (two elements of the lexicon merge and the created unit is again merged with a further element directly extracted from the lexicon) and 3-merge systems (merge two elements created by previous operations of merge). We add a quantitative dimension to the established qualitative dimension discussed in the theory (Rizzi 2016) by investigating large-scale corpora representative of three populations of speakers: adult grammar (102 treebanks/101 languages), typically developing children (2 corpora/English and Chinese) and children with atypical development (1 corpus). The results confirm the predictions in Rizzi (2016): every language in our data set exploits 3-merge systems and less complex systems are the preferred options in early grammars.

Keywords: *Large Datasets, N Merge System, Quantitative Syntax, Subject*

1. Introduction

This paper explores quantitative results based on theoretical assumptions related to the predictions of N-merge systems (Rizzi 2016: 142), systems ranked from minimum to a maximum of complexity in terms of the computational devices and derivational operations they require, in adult grammar and two developmental populations based on large-scale corpora (discussed in details in sub-section 2.2). In particular, we here develop a series of (theory-internal) queries to extract frequencies of 2-merge systems (two elements of the lexicon merge and the created unit is again merged with a further element directly extracted from the lexicon) and 3-merge systems (merge between two elements created by previous operations of merge) in large-scale corpora.

We focus our search on a specific syntactic phenomenon, namely the nature of external arguments (henceforth, EA). EA are assumed to be generated in the verbal domain (vP) of transitive and unergative verbs (Perlmutter 1978; Belletti and Rizzi 1981; Chomsky 1981; Hale and Keyser 1998; for a discussion on externality of arguments; Gallego 2008; Preminger 2008; for more fine-grained cartographic analyses of the vP layer, Ramchand 2008; Si 2019). If no further operation is involved (e.g., smuggling in the case of passivization, Collins 2005; Belletti and Collins 2021), the functional projection hosting EAs in the vP of transitive and unergative verbs are to be analysed as the locus of generation of the "classical" notion of clausal subjects (which will further move to a dedicated criterial position, see Rizzi 2015b: 17ff, for an overview) in these architectures.

EAs represent a first point of investigation for quantitative analyses of N-merge systems since basic layers of complexity arise. Let us compare, for example, the English sentences in (1a) and (1b), where the verbal element *write-* stands for an uninflected form generated within the vP layer.

(1) a. $[_{HEAD}$ She $[_{XP}$ $[_{HEAD}$ write-] $[_{XP}$ the paper]]]
b. $[_{XP}$ [The young linguist] $[_{XP}$ $[_{HEAD}$ write- $[_{XP}$ the paper]]]]

In (1a), a pronominal element *she* merges with an already formed complex phrase composed of *write-* and *the paper*, while in (1b) the EA of the verb is a complex element created by the combination of the definite article *the*, the adjective *young* and the noun *linguist*, which then need to be merged with another complex element *write the paper*. These two configurations bear different layers of difficulty (Rizzi 2015a, 2015b, 2016), as will be discussed throughout the work, and represent two types of N-merge systems: (1a) is a case of 2-merge, while (1b) is a case of 3-merge.

In this study, we provide a quantitative dimension to the established qualitative dimension discussed in Rizzi (2016), investigating large-scale corpora representative of three populations of speakers: adult grammar corpora (102 syntactically annotated treebanks for 101 languages, Universal Dependencies, Nivre 2015; Zeman *et al.* 2020), typically developing (TD) children and children with atypical development datasets (4 morpho-syntactically annotated corpora in Childes; MacWhinney 2000a, 2000b) in order to determine if asymmetries exist between these populations (cf. Borer and Wexler 1987; Guasti 2002, 2017; Kam and Newport 2005; Friedmann, Belletti and Rizzi 2009; Durlemman *et al.* 2015; Stanford 2020, ch.1 for an overview).

We aim to detect if occurrences of generation of complex external arguments (3-merge system, in 1b) can be found in all the adult grammar corpora under investigation and if asymmetries in distributions between 3-merge (1b) and 2-merge (1a) systems arise, in terms of frequencies, in both adult and child grammar. Following Merlo and Stevenson (1998: 134), frequency is intended here as a measure to quantify a role played by grammar (Merlo and Stevenson 1998; Bresnan, Dingare and Manning 2001; Merlo 2016; Samo and Merlo 2019; Gulordava and Merlo 2020). We aim to show that the theoretical predictions on the status of N-merge (Rizzi 2016) quantitatively result in frequencies. This work should represent a first step towards a more complex mapping between theoretical assumption on the complexity of structures and observational data extracted from large datasets.

In section 2, we present a typology of N-merge computational systems (based on Rizzi 2016) which can be naturally ranked from minimum to a maximum of complexity in terms of the computational devices and derivational operations they require. Section 3 presents relevant methods in quantifying complexity with respect to observed frequencies: we verify the

consistency of the predictions of the typology postulated in Rizzi (2016) with the data available from corpus studies in large-scale on a wide range of natural languages syntactically-annotated treebanks and selected corpora from language acquisition. In section 4, we then turn to empirical evidence extracted from syntactically annotated adult treebanks. Once baseline conditions have been established in adult grammar, we investigate two child grammar corpora in English and Chinese, and an English corpus of children with atypical development in section 5. Finally, section 6 discusses and concludes.

2. External arguments and N-merge systems

2.1 Some notes on the cartography of external arguments and subject positions

A wealth of literature has investigated the formal nature of subject and its comparative dimensions (Rizzi 1982; Jaeggli and Safir 1989; Biberauer *et al.* 2010 among others). The subject is considered as an obligatory element (Chomsky 1981) generated as an argument inside the vP (Larson 1988; Borer 1994; Chomsky 1995; Si 2019), which then moves to the Specifier of the Inflectional layer IP/Tense layer TP (Pollock 1989; Koopman and Sportiche 1991; Cardinaletti 2004; Rizzi 2006, 2015a, 2015b; Berthelot 2017) to satisfy subject requirements (e.g., subject criterion in Rizzi 2006; 2015b as a reformulation of classical Chomsky 1981's Extended Projection Principle).

In languages like English, the EA undergoes obligatory movement to a dedicated criterial position SubjP (Cardinaletti 2004; Rizzi 2006) shown in (2).[1] This contrasts with the lack of obligatory movement of the other argument generated in transitive constructions, the 'object of the verb', referred to as the 'internal argument' (IA).

(2) [SpecSubjP The linguist [EA <The linguist> [vP write- [IA the paper]]]]

In transitive verb constructions, if passivization is not involved (in terms of smuggling, Collins 2005) the EA represents the element which will undergo movement to the dedicated functional projection.

Naturally, subjects target other positions if they bear the relevant features. For example, subjects can enter the syntax of cleft structures (Belletti 2015) and relative clauses (Cinque 2014). If the elements bear relevant features (e.g., +Top, +Focus, +Q), subjects can move to dedicated functional projections for topic and focus positions (Frascarelli and Hinterhölzl 2007; Bianchi and Frascarelli 2010; Bianchi, Bocci and Cruschina 2015; Bonan 2019) within both the Left Periphery (Rizzi 1997; Rizzi and Bocci 2017) and the low vP-periphery (Belletti 2004) of the clause. The locus of generation and the landing site of the movement is crucial for explaining asymmetries of subject vs. non-subject configurations in adult processing (Frauenfelder, Segui and Mehler 1980; Chesi and Canal 2019), in TD children (Friedmann, Belletti and Rizzi 2009; Belletti *et al.* 2012), in atypical development (Durrleman *et al.* 2015; Stanford 2020) and in language pathology (e.g., aphasic patients, Grillo 2008; Martini *et al.* 2020; Alzheimer's Disease, Caloi 2013).

[1] A finer cartography of subject positions for different subject elements is provided in Cardinaletti (2004: 116, 154-156).

Linguistic variability can be also detected by the realization of the overt realization of the subject position. Among the syntactic strategies available, languages may drop the subject in certain contexts (Rizzi 1982; Haegeman 1990; Frascarelli 2007; for a typology of languages see Biberauer *et al.* 2010). This will be further discussed in section 4.

2.2 A typology of N-merge systems

Rizzi (2016), rediscussing results of decades of study in generative grammar, takes as starting point the idea that a language system is thought to require at least three main components. The first component is the lexicon, formed by a finite list of items representing linguistic items. The second component is a combinatorial device merge, whose role is to create complex expressions with elements extracted from the lexicon. The operation of "merge" (Chomsky 1957, 1995, 2013; Moro 1997; discussed as a biological primitive in Dehaene *et al.* 2015; see also Boeckx and Theofanopoulou 2014; Murphy 2015), forms a minimal binary tree by combining two elements in order to create a new linguistic object. Finally, a hypothetical (set of) working space(s) is required to apply reiterated operations of merge. Given such components, Rizzi identifies a typology of merge systems with a hierarchy of complexity, "ranked in terms of their generative capacity and of the computational resources they need" (Rizzi 2016: 142).

We here introduce the notion of head [head] and the notion of maximal projection [XP]. Following Rizzi (2015b: 20), we take heads to be directly extracted from the lexicon and maximal projections to be syntactic objects created by merge operations.[2]

Rizzi (2016: 143) proposes a transparent typology of systems labelled according to the number of merge operations: 0-merge, 1-merge, 2-merge and 3-merge systems, with the latter two as objects of our investigation.

In 0-merge systems, an element called head is directly extracted from the lexicon and sent to the systems of sounds and meaning. Only one device is exploited and there are no instances of merge. This system predicts languages based on single word utterances, where head means directly extracted from the lexicon. This system could represent the basis of a subset of linguistic systems, such as those, following Rizzi (2016: 144), belonging to a large set of monkey populations (see Schlenker *et al.* 2016 for an in-depth study) and of early phases of language acquisition (one-word utterances, following Guasti 2002: 24 and citation therein).

1-merge is a system that uses both devices: two elements of the lexicon are merged creating a phrase. Once the phrase is built, it is sent directly to the interface without any recursive procedures. This system could represent an option in a subset of other animals' linguistic systems, such as, following Rizzi (2016 on comment of Schlenker *et al.* 2016), the morphosyntax of a smaller set of monkey populations which produce some rudimentary forms of combinatorial systems.

A 2-merge system is a system in which two elements taken from the lexicon are merged, and a second working space combines a third element from the lexicon. This system is more complex than 0-merge and 1-merge systems, but still relatively simple. Two elements of the lexicon merge and the created unit is again merged with a further element directly extracted from the lexicon. Even if the system seems to be a perfect match between economy of computation and complexity, it does not capture the behaviour of natural languages. Moreover, this system would only lead to either uniformly right or uniformly left tree structures. As expected, such a conclusion might prove to be descriptively inadequate, as predicted by Rizzi (2016: 143).

[2] The discussion on branching direction (left- vs. right- branched languages, see Kayne 1994) and the direction of phrase structure building (bottom-up vs. top-down, see Chesi 2012, 2015) is beyond the scope of this paper.

Finally, 3-merge systems are able to merge two elements created by previous operations of merge. This requires two working systems, lexicon and the operation of merge.

Following the description provided by Rizzi "human languages manifest the full power of 3-merge systems [XP, XP]: no human language is limited to the use of single words (0-merge) or just two-word sequences (1-merge, [head-head]), or to disallow complex specifiers (2-merge, [head, XP]). In other words, human systems possess all the three configurations" (Rizzi 2016: 144): this "human" nature discussed in Rizzi (2016) restricts our field of interest to the last two systems only, 2-merge and 3-merge.

3. Quantifying Hypotheses

A preliminary research question of this paper is directly extracted by Rizzi (2016: 144).

> A 2-merge system would only permit external arguments consisting of one word like *[he [will [meet [the girl]]]]*, but not of two words like *[[the boy] [will [meet [the girl]]]]* (a structure which would require the power of a 3-merge system): no human language appears to have this limitation and disallow complex specifiers. (2016: 144)

The study addresses the prediction in large-scale corpora to investigate whether this conclusion can be made. Following Merlo (2016 and related works), we use corpus counts in the spirit of the computational quantitative syntax framework, observing differentials in frequencies as the expression of underlying grammatical properties. Frequency is here adopted as a measure to quantify linguistic proposals (on the role of frequency in grammar, acquisition and formal rules; Yang 2013, 2015; see Ibbotson 2013 for a usage-based grammar account) and represents a dependent variable to test linguistic models (Merlo 2016 and related works).

The first research question then investigates a linguistic evidence from a large set of languages, expecting that in no language is the 3-merge system absent. In other words, every language should have at least one occurrence of an external argument built with a 3-merge system. We state the research question in H_1.

H_1: The frequency of 3-merge systems in adult grammar corpora should be $\neq 0$ in every language under investigation.

Rizzi (2016) does not predict any preference, but the two systems should be equally exploited by adult grammars. In this paper, we contribute with a quantitative analysis of these two elements to observe whether we can observe a preference for 3-merge systems over 2-merge systems or vice versa.

Our second research question involves a dimension in terms of development in acquisition. Data from corpora and experimental studies in language acquisition clearly show that 2- and 3-merge systems co-exist at early stages of child grammar (Guasti 2002, 2017 and Belletti and Guasti 2015, for an overview; Friedmann, Belletti and Rizzi 2020). However, the exploitation of a different number of computational devices (2-merge: lexicon, merge, one working space; 3-merge: lexicon, merge, two working spaces) results in asymmetries. Along the same lines, we expect that, if there is any (non-marginal) increase in the production, it should affect 2-merge systems earlier than 3 merge systems.

H_2: The distributions of emergence of 2- and 3-merge systems in child grammar should differ.

Finally, we also investigate a corpus of atypical development to observe patterns whether similar asymmetries can be found.

In order to quantitatively answer these questions, we extract data from large-scale resources. To facilitate our search, we rely on large (morpho-)syntactically annotated corpora. The materials and the methods in retrieving the frequencies of the two systems are presented in the relevant sections and sub-sections.

Following Merlo (2016) and related works, our quantitative hypotheses presented here are to be contrasted to a H_0 hypothesis that would predict that grammatical properties are uncorrelated to frequencies.

To answer H_1, we investigate corpus evidence from adult grammar presented in section 4. As for H_2, we extract our data from child spontaneous production repositories presented in section 5.

4. A crosslinguistic study in Adult Grammar

The size of linguistic materials on adult grammar are huge and heterogeneous, therefore we establish a set of parameters in choosing the appropriate material for our research. First of all, our goal is to automatically gather as much linguistic evidence as possible. In order to fully automatize our search and make it replicable at different layers, we chose (morpho)-syntactically annotated corpora. A fundamental annotation is in terms of the grammatical (syntactic) function in the sentence, to detect occurrences of external arguments (subjects in most syntactic annotations' tools). If the grammatical functions are possibly combined with a Part-of-Speech tag (henceforth, PoS) annotation, it is faster to classify external arguments (subjects) as maximal projections (noun, proper nouns) and heads (pronominal entities). Merging these two elements provides the right amount of information to detect and differentiate the two merge systems under investigation. Finally, we used a syntactically annotated database allowing us to investigate as many languages (and language families) as possible. Candidates are large, multilingual, homogeneously-annotated data sets, as the treebanks provided by the Universal Dependencies (Nivre 2015; Zeman *et al.* 2020), which will be briefly described in sub-section 4.1.

4.1 Materials and Methods

Our material is extracted from syntactically annotated treebanks following the guidelines of the Universal Dependencies (henceforth UD, version 2.7, Zeman *et al.* 2020) annotation scheme, allowing direct comparison across languages. We take into consideration 102 treebanks for 101 languages. To factor out problems related to genre classification of the treebanks, we chose, when possible, the biggest and the most heterogeneous treebank for each language. A parameter in preferring a specific treebank was given by the number and the quality of types of registers provided by UD guideline screen.[3] We avoided treebanks with less than 50 trees and, when possible, parallel treebanks (Ahrenberg 2007; Volk, Graën and Callegaro 2015). Detailed information of the materials (size, references) is provided in Table 1 together with the results of study 1.

[3] <http://www.universaldependencies.org > (06/2021).

For Italian, we adopted two treebanks to detect, if any, genre effects on the distribution of complex structures (in the spirit of Samo, Zhao, and Gamhewage 2020 and Zhao *et al.* 2021): the ISDT treebank v.2.7 (Bosco, Montemagni and Simi 2013, text genres: legal, news, wiki) and TWITTIRO treebank 2.7 (Cignarella *et al.* 2018, text genres: social media).

Beyond investigating the presence of 3-merge systems in 97 monolingual treebanks, we also controlled for specific populations: a sign language treebank (Swedish Sign Language), two bilingual treebanks (Hindi-English and Turkish-German), and two treebanks of learner essays of L2 (English and Chinese L2). For recent generative (cartographic) analysis on these populations, see Bross (2020) for Sign Languages; Shim (2016) for code-switching; and Di Domenico, Baroncini and Capotorti (2020) for L2, and references therein.

We performed two studies. In the first study, we observed the distribution of 3-merge systems only. In the second study, we aimed to quantify whether there is a preference between the two systems. In this latter study, we also took into consideration the role of null subject, following typological classifications provided by the World Atlas of Language Structures (WALS, Dryer and Haspelmath 2013), to evaluate our data in light of the availability of choices in a given language.[4] We considered non-null subject languages those labelled as "Obligatory pronouns in subject position" in Dryer (2013) and WALS as null subject languages, whereas the set of labels ("Subject clitics on variable host", "Subject affixes on verb", "Optional pronouns in subject position" and "Mixed") as languages allowing null subject constructions, but we do not discuss the different fine-grained differences among the different groups (see, for example, Neeleman and Szendrői 2007; Biberauer *et al.* 2010; Holmberg and Roberts 2013; Frascarelli and Casentini 2019). We analysed 48 treebanks for 47 languages, since both treebanks of Italian are investigated to detect, if any, further effects of text genre.

Finally, to detect cross-linguistic genre effects, we ran a linear regression (inspired by Merlo and Ouwayda 2018) and automatically ranked the costs of specific genres in the distribution of the treebank. We used the Waikato Environment for Knowledge Analysis, WEKA v.3.8.2 (Hall *et al.* 2009) to derive the best linear regression model of this data. Each treebank is encoded as a vectorial representation and every genre is encoded as an indicator variable: 0 and 1 indicates if the genre is present or not in the relevant treebank. Positive and negative coefficients indicate the difference from a predicted frequency: genre that are associated with lower distributions will have negative coefficients, and those associated with higher distributions will get positive coefficients. We used a leave-one-out cross-validation.

All the data were extracted with the online web-based tool Match Grew.[5] We followed the guidelines for mapping UD into cartographic representation (and viceversa) proposed in Samo (2019).

Among the different syntactic labels, provided by the annotation scheme, we focused only on the core dependency subject, represented as *nsubj*: as for the modifiers of the subject, we looked at the dependencies starting from the subjects provided by *det* 'determiner', *case* 'case marking' and *amod* representing any nominal modifiers. Another important ingredient is the annotation of the PoS tag: we were therefore able to detect the pronominal or full nominal nature of the external argument.

The query providing 3-merge occurrences provided the occurrence of at least one complex specifier (modifiers) of subjects (3-merge systems) in bi-argumental/transitive constructions (the verb governs a subject and an object dependency). The query detected all of the occurrences

[4] <https://wals.info/feature/101A#2/18.0/148.2> (06/2021).

[5] <www.grew-match.fr> (06/2021)

of sentences given a variable being dependent of the dependency *nsubj* (subject) and governor of one of the set of dependencies *det* (determiner), *case* (case marking), *amod* (modifier such as adjectives). To this count, we added the frequencies of subjects annotated as PROPN (proper nouns) as PoS in transitive constructions. Proper nouns have been considered XP, following theoretical considerations in Longobardi (1994: 641 and related reference), since the nominal element is described as targeting a D position within the DP. Naturally, the utterances retrieved may have a higher layer of complexity, involving the presence of other arguments (e.g., indirect objects), adverbials and complements.

For 2-merge, we developed a query which retrieved dependents on a subject dependency, whose part-of-speech is a pronoun or bare nominals and which do not govern any modification (*det, case, amod*) dependency.

The queries for the adopted tool and a naturally occurring sentence for type extracted from Italian (ISDT treebank, Bosco, Montemagni and Simi 2013), are given in (3).[6]

(3) Queries (grewmatch.fr, accessed, 05.06.2021) and naturally occurring examples in Italian

a. 3-merge systems for subjects

i. pattern {a -[nsubj]-> b; b -[det|case|amod]-> c; b [upos = "NOUN"]; a-[obj]->d}

ex. i pompieri hanno isolato la sala. (ISDT-isst-tanl-7)
the firefighters have seal.off the room
'Firefighters sealed off the room'

ii. pattern {a -[nsubj]-> b; b [upos = "PROPN"]; a-[obj]->c}

ex. Moretti rappresenta il cinema di oggi (ISDT-isst-tanl-1511)
Moretti represents the cinema of today
'Moretti represents today's cinema'

b. 2-merge systems for subjects

i. pattern {a -[nsubj]-> b; b [upos = "PRON"]; a-[obj]->d} without {b -[det|case|amod]-> c}

ex. Io studio l' inglese (ISDT-isst-tanl-2961)
I study the English
'I study English'

ii. pattern {a -[nsubj]-> b; b [upos = "NOUN"]; a-[obj]->d } without {b -[det|case|amod]-> c }

ex. Graffiti imbrattano le città. (ISDT-isst-tanl-1924)
Graffiti litter the cities
'Graffiti litter cities'

Results are discussed in sub-section 4.3.

4.3 Results

Table 1 summarizes the results of study 1 confirming H1 (The frequency of 3-merge systems in adult grammar corpora should be ≠ 0 in every language under investigation). As Table 1 shows, every language under investigation, even in smaller treebanks, exploits 3-merge

[6] The tool adopted in the investigation (accessed February 18th, 2021) provided only the first 1000 occurrences of the query, a coefficient has been calculated on the basis of the occurrences. This coefficient is calculated to provide a better understanding of a predictive tool. The trees are used as coefficients instead of subjects to keep an analysis in terms. Being *I* an imputed count, *F* the frequency of the result, and *C* the percentage of the exploitation of the corpus, the imputed count is derived from the formula. $I = F / C$.

systems (see raw frequencies in column FREQ(UENCY) 3-MERGE and the distribution among trees in column FREQ(UENCY)/TREES). Table 1 shows the results for 102 treebanks, confirming H_1.

LANGUAGE	TREEBANKS GENRE	TREES	TOKENS	FREQ. 3-MERGE	FREQ/ TREES	References (if website, https://universaldependencies.org/treebanks/[...])
Afrikaans	AfriBooms L, NF	1934	51210	445	0.23	[…]/af_afribooms/index.html
Akkadian	RIAO NF	1799	23701	132	0.07	Luukko *et al.* (2020)
Albanian	TSA W	60	982	7	0.12	[…]/sq_tsa/index.html
Amharic	ATT B, F, GE, N, NF	1074	11084	121	0.11	Ephrem Seyoum, Miyao and Yimam (2018)
Ancient Greek	PROIEL B, NF	17080	231079	1937	0.11	Haug and Jøhndal. (2008)
Apurina	UFPA N, NF	75	635	7	0.09	Freitas (2017)
Arabic	NYUAD N	19738	758627	12466	0.63	[…]/ar_nyuad/index.html
Mod. East. Armenian	ArmTDP BL, F, GE, L, N, NF	2502	55132	244	0.10	Yavrumyan (2019)
Mod. Stand. Assyrian	AS N, NF	57	510	3	0.05	[…]/aii_as/index.html
Bambara	CRB N, NF	1026	14849	34	0.03	[…]/bm_crb/index.html
Basque	BDT N	8993	130436	1153	0.13	[…]/eu_bdt/index.html
Belarusan	HSE F, L, N, NF, P, SM	23534	298867	1283	0.05	[…]/be_hse/index.html
Bhojpuri	BHTB N, NF	357	7022	29	0.08	Ojha and Zeman (2020)
Breton	KEB F, GE, N, NF, P, W	888	10942	88	0.10	Tyers and Ravishankar (2018)
Bulgarian	BTB F, L, N	11138	167287	1327	0.12	[…]/bg_btb/index.html
Buryat	BDT F, GE, N	927	11112	49	0.05	Badmaeva and Tyers (2017)
Cantonese	HK S	1004	14922	31	0.03	[…]/yue_hk/index.html
Catalan	AnCora N	16678	548649	18088	1.08	[…]/ca_ancora/index.html

Chinese	GSD [W]	4997	128288	954	0.19	[...]/zh_gsdsimp/index.html
Chinese - L2	CFL [LE]	451	7707	26	0.06	[...]/zh_cfl/index.html
Chukchi	HSE [S]	1004	7211	14	0.01	[...]/ckt_hse/index.html
Classical Chinese	Kyoto [NF]	48434	281556	3990	0.08	[...]/lzh_kyoto/index.html
Sahidic Coptic	Scriptorium [B, F, NF]	1873	50504	153	0.08	[...]/cop_scriptorium/index.html
Croatian	SET [N, WEB, W]	9010	208419	2222	0.25	Agić and Ljubešić (2015)
Czech	PDT [N, NF, R]	87913	1596965	12808	0.15	Bejček *et al.* (2014)
Danish	DDT [F, N, NF, S]	5512	106245	954	0.17	Johannsen, Alonso and Plank (2015)
Dutch	Alpino [N]	13578	222179	3146	0.23	[...]/nl_alpino/index.html
English	EWT [BL, EM, R, SM]	16622	271478	1681	0.10	[...]/en_ewt/index.html
English-L2	ESL [LE]	5124	102805	339	0.07	[...]/en_esl/index.html
Erzya	JR [F]	1690	18838	107	0.06	Rueter and Tyers (2017)
Estonian	EDT [A, F, N, NF]	30972	469143	2441	0.08	[...]/et_edt/index.html
Faroese	OFT [W]	1208	11210	24	0.02	Tyers *et al.* (2018)
Finnish	FTB [GE]	18723	178335	952	0.05	[...]/fi_ftb/index.html
French	GSD [BL, N, R, W]	16341	416740	3865	0.24	[...]/fr_gsd/index.html
Galician	CTG [L, M, N, NF]	3993	142830	2562	0.64	[...]/gl_ctg/index.html
German	HDT [N, NF, R, WEB, W]	189928	3589318	85216	0.45	Borges Völker *et al.* (2019)
Gothic	PROIEL [B]	5401	60737	240	0.04	[...]/got_proiel/index.html
Greek	GDT [N, S, W]	2521	65962	931	0.37	Prokopidis and Papageorgiou (2017)
Hebrew	HTB [N]	6216	167633	522	0.08	Tsarfaty (2013)
Hindi	HDTB [N]	16647	368351	5930	0.36	Bhat *et al.* (2017)

Hindi-English	HIENCS [SM]	1898	28807	185	0.10	Bhat *et al.* (2018)
Hungarian	Szeged [N]	1800	43832	517	0.29	[…]/hu_szeged/index.htm[l]
Icelandic	PUD [N, W]	1000	19833	171	0.17	[…]/is_pud/index.htm[l]
Indonesian	GSD [BL, N]	5593	127516	1451	0.26	[…]/id_gsd/index.html
Irish	IDT [F, L, N, W]	4910	120879	1044	0.21	[…]/ga_idt/index.html
Italian	ISDT [L, N, W]	14167	312547	3166	0.22	Bosco, Montemagni and Simi (2013)
Italian Twitter	TWITTIRO [SM]	1424	31029	295	0.21	Cignarella *et al.* (2018)
Japanese	GSD [BL, N]	8071	200676	1135	0.14	[…]/ja_gsd/index.html
Karelian	KKPP [N, NF, WEB]	228	3322	10	0.04	[…]/krl_kkpp/index.html
Kazakh	KTB [F, N, W]	1078	11614	63	0.06	Makazhanov *et al.* (2015)
Komi Permyak	UH [F]	81	920	2	0.02	Reuter, Partanen and Ponomareva (2020)
Komi Zyrian	Lattice [F]	435	5437	14	0.03	Partanen *et al.* (2020)
Korean	Kaist [A, F, N]	27363	377453	523	0.02	[…]/ko_kaist/index.html
Kurmanji	MG [F, N]	754	11014	39	0.05	Gökırmak and Tyers (2017)
Medieval Latin	ITTB [NF]	26977	477492	1731	0.06	Cecchini *et al.* (2018)
Latin	PROIEL [B, NF]	18411	218574	877	0.05	[…]/la_proiel/index.html
Latvian	LVTB [A, F, L, N, S]	13643	234179	1186	0.09	[…]/lv_lvtb/index.html
Lithuanian	ALKSNIS [F, L, N, NF]	3642	73693	153	0.04	Bielinskienė *et al.* (2016)
Livvi	KKPP [N, NF, WEB]	125	1757	6	0.05	[…]/olo_kkpp/index.htm[l]
Maltese	MUDT [F, L, N, NF, W]	2074	46236	352	0.17	[…]/mt_mudt/index.html
Manx Gaelic	Cadhan [B, BL, F, N, NF, SM, WEB, W]	291	6445	17	0.06	[…]/gv_cadhan/index.htm[l]

Marathi	UFAL [F, W]	466	4315	16	0.03	[...]/mr_ufal/index.htm[l]
Mbya Guarani	Thomas [NF]	98	1416	2	0.02	[...]/gun_thomas/index.htm[l]
Moksha	JR [N, NF]	167	1681	3	0.02	Reuter (2018)
Munduruku	TuDeT [N, NF]	62	333	1	0.02	[...]/myu_tudet/index.ht[ml]
Naija	NSC [S]	9242	149971	502	0.05	[...]/pcm_nsc/index.html
North Sami	Giella [N, NF]	3122	29967	151	0.05	Tyers and Sheyanova (2017)
Norwegian Bokmål	Bokmaal [BL, N, NF]	20044	330265	1584	0.08	[...]/no_bokmaal/index.htm[l]
Norwegian Nynorsk	Nynorsk [BL, N, NF]	17575	318928	2342	0.13	Velldal, Øvrelid and Hohle (2017)
Old Church Slavonic	PROIEL [B]	6338	63901	231	0.04	[...]/cu_proiel/index.html
Old French	SRCMF [L, NF, P]	17678	188418	1914	0.11	Stein and Prévost (2013)
Old Russian	TOROT [L, NF]	16944	166724	827	0.05	Eckhoff and Berdičevskis (2015)
Persian	Seraji[F, L, M, N, NF, S, SM]	5997	158917	435	0.07	[...]/fa_seraji/index.html
Polish	PDB [F, N, NF]	22152	372188	1539	0.07	[...]/pl_pdb/index.html
Portoguese	GSD [BL, N]	12078	331931	3938	0.33	[...]/pt_gsd/index.html
Romanian	RRT [A, F, L, M, N, NF, W]	9524	228035	1138	0.12	[...]/ro_rrt/index.html
Russian	SynTagRus___[F, N, NF]	61889	1169630	4990	0.08	Droganova, Lyashevskaya and Zeman(2018)
Sanskrit	Vedic [NF]	3997	31114	71	0.02	Hellwig *et al.* (2020)
Scottish Gaelic	ARCOSG [F, N, NF, S]	3173	63590	131	0.04	Batchelor (2019)
Serbian	SET [N]	4384	102057	1045	0.24	[...]/sr_set/index.html
Skolt Sami	Giellagas [N, NF, S]	104	1456	4	0.04	[...]/sms_giellagas/index.html
Slovak	SNK [F, N, NF]	10604	116734	1072	0.10	Zeman (2017)
Slovenian	SSJ [F, N, NF]	8000	148670	1253	0.16	[...]/sl_ssj/index.html

Spanish	AnCora [N]	17680	567429	11633	0.66	[...]/es_ancora/index.html
Swedish	Talbanken [N, NF]	6026	102884	706	0.12	[...]/sv_talbanken/index.html
Swedish Sign	SSLC [S]	203	1813	6	0.03	[...]/swl_sslc/index.html
Swiss German	UZH [BL, F, N, NF, W]	100	1544	17	0.17	Aepli and Clematide (2018)
Tagalog	Ugnayan [F, NF]	94	1191	19	0.20	[...]/tl_ugnayan/index.html
Tamil	TTB [N]	600	10181	73	0.12	[...]/ta_ttb/index.html
Telugu	MTG [GE]	1328	7739	57	0.04	[...]/te_mtg/index.html
Thai	PUD [N, W]	1000	23322	139	0.14	[...]/th_pud/index.html
Turkish	BOUN [N, NF]	9761	132144	649	0.07	Türk *et al.* (2020)
Turkish-German	SAGT [S]	1891	33837	51	0.03	Çetinoğlu and Çöltekin (2019)
Ukrainian	IU [BL, EM, F, GE, L, R, SM, WEB, W]	7060	129384	713	0.10	[...]/uk_iu/index.html
Upper Sorbian	UFAL [NF, W]	646	11842	94	0.15	[...]/hsb_ufal/index.html
Urdu	UDTB [N]	5130	143207	2199	0.43	Bhat *et al.* (2017)
Uyghur	UDT [F]	3456	43962	160	0.05	[...]/ug_udt/index.htm[l]
Vietnamese	VTB [N]	3000	46754	514	0.17	[...]/vi_vtb/index.htm[l]
Warlpiri	UFAL [GE]	55	369	4	0.07	[...]/wbp_ufal/index.html
Welsh	CCG [F, GE, N, NF, W]	1657	34568	34	0.02	Heinecke and Tyers (2019)
Wolof	WTB [B, W]	148	46365	319	2.16	[...]/wo_wtb/index.html
Yoruba	YTB [B, W]	318	8561	67	0.21	[...]/yo_ytb/index.html

Table 1 – Treebank, Size (trees and tokens), raw frequencies of 3-merge and distribution of 3-merge out of number of trees in every language under investigation. Genre abbreviations: A = Academic, B = Bible, BL = Blog, EM = Emails, F = Fiction, G = Government, GE = Grammar Examples, L = Legal, LE = Learner-essays, M = Medical, N = News, NF = Nonfiction, P = Poetry, S = Spoken, SM = Social Media, R = Reviews, W= Wikipedia, WEB = Web

The distribution of 3-merge with respect to the trees on the treebank does not correlate with the size of the treebanks (r = 0.18). As for text genre in Italian, the results show minimal differences between the distribution of 3-merge systems in the two investigated treebanks (ISDT 0.22, Twittiro 0.21). The Linear Regression model's results (r = 0.03; Academic +0.3268, Web +0.1076, Fiction -0.146), shows a marginal increase of 3-merge in Academic and Web extracted material, whereas fiction (e.g., novels) seems to have a negative impact on the production of 3-merge systems. Future studies should investigate dimensions of variation among genres.

The second study investigated whether there is a preference between the two systems (2-merge and 3-merge). Preference is intended here in frequencies. Table 2 summarizes the results.[7]

Language	Null (p)	Freq3m	%3m	Freq2m	%2m	Pref. (k)	*Obj (N)*	*z-test*
Albanian	Yes	7	0.37	12	0.63	2merge	*34*	p = .459329.
Amharic	Yes	121	0.24	377	0.76	2merge	572	p < .000001.
Arabic	Yes	12466	0.51	11998	0.49	3merge	118666	p < .000001.
Armenian	Yes	244	0.33	506	0.67	2merge	2338	p < .000001.
Basque	Yes	1153	0.51	1114	0.49	3merge	7522	p < .000001.
Breton	Yes	88	0.77	27	0.23	3merge	393	*p* = .000005.
Bulgarian	Yes	1327	0.40	2000	0.60	2merge	6271	*p* = .032073.
Catalan	Yes	18088	0.76	5651	0.24	3merge	37089	*p* < .000001.
Chinese	Yes	954	0.33	1934	0.67	2merge	7748	*p* < .000001.
Croatian	Yes	2221.5	0.57	1685	0.43	3merge	8796	*p* < .000001.
Czech	Yes	12808	0.55	10391	0.45	3merge	54265	*p* < .000001.
Danish	No	954	0.27	2569	0.73	2merge	5011	*p* < .000001.
Dutch	No	3146.4	0.51	3042	0.49	3merge	6956	*p* < .000001.
English	No	1681	0.28	4264	0.72	2merge	12600	*p* < .000001.
Erzya	Yes	107	0.42	148	0.58	2merge	832	*p* < .000001.
Estonian	Yes	2440.9	0.27	6526	0.73	2merge	21290	*p* < .000001.
Finnish	Yes	952	0.29	2358	0.71	2merge	8661	*p* < .000001.
French	No	3864.5	0.51	3750	0.49	3merge	12708	*p* < .000001.
German	No	85216	0.66	44197	0.34	3merge	154653	*p* < .000001.
Greek	Yes	931	0.80	230	0.20	3merge	2382	*p* < .000001.
Hebrew	Yes	522	0.34	1036	0.66	2merge	3925	*p* < .000001.
Hungarian	Yes	517	0.61	334	0.39	3merge	1763	*p* = .000564.

[7] The binomial test gives us the probability of k successes (the number of the preferred merge system) in N independent trials (the number of objects, therefore transitive constructions), given a base probability p (the probability given by the null subject nature of the language) of an event. The binomial test (*z-test*) gives us the (one-tailed) probability of exactly the observed counts.

Indonesian	No	1451.1	0.50	1434	0.50	3merge	5795	$p < .000001$.
Irish	Yes	1043.8	0.62	648	0.38	3merge	4561	$p < .000001$.
Italian	Yes	3165.9	0.69	1444	0.31	3merge	10239	$p = .000004$
Italian Twitter	Yes	295	0.74	103	0.26	3merge	1222	$p < .000001$.
Korean	Yes	523	0.26	1501	0.74	2merge	23605	$p < .000001$.
Latvian	No	1186	0.29	2858	0.71	2merge	9780	$p < .000001$.
Lithuanian	Yes	153	0.25	464	0.75	2merge	2505	$p < .000001$.
Mbya Guarani	Yes	2	0.14	12	0.86	2merge	63	$p = .013168$.
Norwegian Bokmål	No	2591.8	0.28	6712	0.72	2merge	13904	$p = .000024$.
Norwegian Nynorsk	No	2342.3	0.24	7453	0.76	2merge	13218	$p < .000001$.
Persian	Yes	435	0.24	1393	0.76	2merge	3870	$p = .00004$.
Polish	Yes	1539	0.37	2596	0.63	2merge	15273	$p < .000001$.
Portoguese	Yes	3938.2	0.67	1914	0.33	3merge	11074	$p < .000001$.
Russian	No	4990.4	0.37	8532	0.63	2merge	33928	$p < .000001$.
Scottish Gaelic	No	131	0.18	608	0.82	2merge	1872	$p < .000001$.
Serbian	Yes	1045	0.59	738	0.41	3merge	3441	$p = .000549$.
Slovenian	Yes	1253	0.53	1114	0.47	3merge	7306	$p < .000001$.
Spanish	Yes	11633	0.73	4354	0.27	3merge	32061	$p < .000001$.
Swedish	No	706	0.22	2479	0.78	2merge	4241	$p < .000001$.
Thai	Yes	139	0.33	285	0.67	2merge	1734	$p < .000001$.
Turkish	Yes	649	0.39	1000	0.61	2merge	7402	$p < .000001$.
Upper Sorbian	Yes	94	0.57	71	0.43	3merge	368	$p = .00141$.
Uyghur	No	160	0.20	646	0.80	2merge	2301	$p < .000001$.
Vietnamese	Yes	514	0.33	1053	0.67	2merge	4078	$p < .000001$.
Warlpiri	Yes	4	0.09	41	0.91	2merge	50	$p < .000001$.
Wolof	Yes	319	0.18	1414	0.82	2merge	3319	$p < .000001$.
Yoruba	Yes	67	0.17	332	0.83	2merge	536	$p < .000001$.
Total		**190181**	**0.55**	**155351**	**0.45**	**3merge**	**692221**	$\boldsymbol{p} < .000001$*.

Table 2 – Language, whether the language is a null subject language establishing *p* (NULL P), frequency and distribution of 3-merge systems (FREQ3M, %3M), frequency and distribution of 2-merge systems (FREQ2M, %2M), preferred system representing the number of observation (K) for the binomial test, number of objects (OBJ) in the treebank representing the number of events (N) and the Z-TEST. * indicates that *p* has the same value for both 2- and 3-merge systems

There are no typological trends (e.g., German and Dutch prefer 3-merge systems, while Norwegian Bokmål and Norwegian Nynorsk show preference for 2-merge systems) and there is no correlation with the size of the corpus (r = 0.20). Considering all the languages as performance of a unique language, we can observe a marginal preference of 3-merge systems. This result is not indicative, since the distribution varies among languages having specific parametric values (null-subject vs. non-null-subject languages). Similarly, the two treebanks of Italian show relatively marginal different distributions (ISDT 0.69, TWITTIRO 0.74), but the linear regression model (r = 0.29, Poetry + 0.4303, Social Media + 0.3888, News +0.1957, Reviews +0.1667, Academic -0.2024, Nonfiction -0.2133, Medical -0.2899, Emails -0.6252) shows that there might be a trend. Indeed, registers like emails and (partially) academic/medical show a reduction of the usage of 3-merge systems. We leave the in-depth investigation of this result to further studies.[8]

The observed results in adult grammar open the question whether developmental paths in acquisition might show interesting dimensions of variation. Therefore, in section 5, we investigate child grammar (typical development and atypical development) to examine whether the complexity of the developing system gives rise to asymmetries.

5. Child Grammar and N-merge: some preliminary results

This section focuses on the frequency of 2-merge and 3-merge systems in typical development from observations in selected corpora of child grammar from Childes (MacWhinney 2000). In this study, we do not make any specific assumption on the structural configuration of child grammar with respect to early production of multiword utterances (see Guasti 2002: chapter 4 and reference therein). For the scope of our paper, we limit our analysis to the mere discussion of quantitative results of the extracted linguistic data adopting the model developed in section 4.

We restricted the investigation to two languages from two unrelated language families, namely English and Chinese. Following Guasti (2002: 101-310), we narrow our search on the period of time between two and four years (and around these two extremes).

5.1 Materials & Methods

After a manual analysis of the corpora collecting data from the languages under investigation, we decided to analyse longitudinal corpora from English and Mandarin Chinese. We investigated data extracted automatically from the *childesdb* (Sanchez *et al.* 2019). We performed the task on R (R development team 2016), isolating only target children's utterances in the relevant age (in terms of months). We only selected those sentences that are annotated for PoS.[9] A manual analysis was also conducted to evaluate the quality of the retrieved data. Some information on the size of the corpora, age range, the number of annotated utterances and references of the corpora under investigation are given in Table 3.

[8] These results are in line with the findings in Samo, Zhao and Gamhewage (2020) and Zhao *et al.* (2021), who have shown that syntactic complexity is cross-linguistically minimized in certain contexts, such as learning contents in public health with respect to social media, encyclopaedic entries and news.

[9] The morpho-syntactic annotation for corpora in Childes tendentially follows the guidelines provided by the annotation schemata MOR (doi:10.21415/T5B97X).

Language	Corpus	Age Range	Children	Annotated Utterances	References
English	Wells	1;6 – 5;0	32	17,964	Wells (1981)
Mandarin Chinese	Tong[A], Zhou3[B]	1;7-3;4[A] 0 :8 – 4 :5[B]	2	14,860	Deng and Yip (2015, 2018)[A]; Zhang and Zhou (2009)[B]

Table 3 – Relevant info on the corpora, age range, number of children, annotated utterances and related reference for the corpora under investigation

We translate the queries developed in section 4 according to the relevant annotation schemata. Both queries are based on the occurrences of patterns of labels in transitive construction given by a verbal element followed by an object. The morpho-syntactically annotated elements considered as 2-merge are personal pronouns (*pro:per, pro:sub*) and bare nominals (*n*); the combination of nominal elements with adjective (*adj*), articles (*det:art),* numerals *(num*) and classifiers (*cl*) were considered 3-merge. As discussed in section 4.1., we considered proper nouns (*n:prop*) as 3-merge elements. A manual analysis has been carried out to evaluate our semi-automatic retrieval. We summarize the queries in (4).

(4) 2-merge {pro:per}/{pro:sub}/{n} + transitive construction

3-merge {{det:art}/ {adj}/ {num} + {cl} + n} / {n:prop} + transitive construction

Naturally occurring sentences like (5a, b) will be labelled as 2-merge, while sentences like (5c) and (5d) as a 3 merge, some examples can be found in (5).

(5) a. English, 2-merge
I want my money (Elspeth, 2;6, ID: 9789443)

b. English, 3-merge
The dog have that ball (Abigail, 3:3, ID: 9761820)

c. Mandarin Chinese, 2-merge
我想画 个衣服 (Xue'er 1;11, ID: 5265930)

wo3	xiang3	hua4	ge4	yi1fu2
I	want	draw	cl	dress

d. Mandarin Chinese, 3-merge
这两个小朋友是男生 (Xue'er, 4;4 , ID: 5259789)

zhe4liang3ge4	xiao3peng2you3	shi4	nan2sheng1
this.two,cl	little-friends	are	boys

As for age of the target child (in months), we grouped the utterances in class intervals of one year (younger than 24 months, from 24 to 36 months, from 36 to 48 months, older than 48 months). Sub-section 5.2 summarizes the results.

5.2 Results

Results confirm H_2 (*the distributions of emergence of 2- and 3-merge systems in child grammar should differ*). Table 4 summarizes the results.

Age	EnUtt	En2mF	En%2m	En3mF	En%3m	ZhUtt	Zh2mF	Zh%2m	Zh3mF	Zh%3m
< 24	1244	20	0.016	2	0.002	2086	39	0.019	8	0.004
24 – 36	7119	551	0.077	52	0.007	6618	338	0.051	38	0.006
36 – 48	7069	783	0.111	79	0.011	4201	260	0.062	27	0.006
> 48	2532	352	0.139	30	0.012	1955	180	0.092	19	0.010

Table 4 – Age (in month), number of utterances in English and in Chinese (EnUtt, ZhUtt), frequency (F) of 2-merge and 3-merge in English and in Chinese (En2mF, En3mF, Zh2mF, Zh3mF) and distributions (%) of 2-merge and 3-merge in English and Chinese (En%2m, En%3m. Zh%2m, Zh%3m)

Our dataset shows that 2-merge system is the preferred option, in terms of distributions, in every age interval and in both languages. Similar cross-linguistic distributions can be observed before 24 months for both 2-merge (0.016 in English, 0.019 in Chinese) and 3-merge (0.002 in English and 0.004 in Chinese). After 48 months, we detect a comparably similar distribution for 3-merge configuration between English (0.012) and Chinese (0.010). The increase of the usage of these structures correlates with the age intervals: we can detect a correlation between age and the distribution of 2-merge ($r = 0.98$, $p < .05$) and 3-merge ($r = 0.97$, $p < .05$) in English. Asymmetries between the two systems are found Chinese: a slightly stronger correlation between age intervals is observed in 2-merge systems ($r = 0.99$, $p < .05$) compared to 3-merge systems ($r = 0.92$, $p = .08$).

Further research will involve a higher data set of languages and utterances (beyond longitudinal corpora) to detect more fine-grained differences.

5.3 Some notes on grammar in atypical development: a manual investigation

The last research question investigates grammars in children with atypical development. We extracted our data from the "Conti 2" corpus (Conti-Ramsden and Dykins 1991; Conti-Ramsden, Hutcheson and Grove 1995; Conti-Ramsden and Jones 1997), which contains transcripts, among others, from three children with specific language impairment (SLI). We first isolated the utterances of the relevant children (Andrew, Colin and Mark), then a manual investigation was carried out, due to the limited size of the dataset. This allowed us to detect every transitive construction and the nature of subjects, including null subjects (NS). Age intervals are different than the previous study in section 5.1. and 5.2., due to the nature of our dataset. We investigate utterances produced by the target children before 4 years, during the interval between 4 and 5 years and between 5 and 6 years, and after 6 years of age. Results are summarized in Table 5.

Age	Utterances	Verbs	Transitive	NS	Frequency 2m	Frequency 3m
<48	66	4	3	3	-	-
48-60	166	38	13	7	6	-

60-72	435	40	17	7	10	-
> 72	651	108	55	6	47	2

Table 5 – Age intervals in months, number of utterances, number of verbs, number of transitive constructions, number of null subjects (NS), frequency of 2-merge systems (2M) and frequency of 3-merge systems

In our (reduced) sample, we can observe that there is a strong preference for 2-merge systems in every age interval. As table 5 shows, 3-merge systems emerge later than 2-merge systems in our dataset. The two naturally occurring utterances are given in (6).

(6) a. a burglar he got it (Andrew, 78 months, ID: 14912223)
b. this one no beep the horn (Colin, 93 months, ID: 14916479)

From the transcriptions at our disposal, we can observe that (6a) might be a case of a topicalized subject or a configuration displaying a hanging topic. The example in (6b) involve a more complicated configuration with lack of agreement. Further research on bigger datasets is welcome to understand the dynamics concerning the relevant populations of speakers.

6. Conclusion

The results provided here shed light on the distribution of the typology of N-merge systems introduced in Rizzi (2016) adopting frequency as a dependent variable to test linguistic proposals (Merlo 2016 and related works).

We observed that in a rich dataset of 101 languages (102 treebanks), we can retrieve at least one occurrence of 3-merge in every variety. Such a result might confirm, at least for the sample we investigated, the prediction on the status of 3-merge systems in natural languages proposed in Rizzi (2016: 144).

Another important research question investigated focussed on preferences for 2-merge and 3-merge systems in adult grammar. We analysed 48 treebanks for 47 languages and we did not observe a clear typological trend (either null subject languages or language families). However, variability in terms of registers seems to be playing a role.

Developmental trends can be analysed via the exploration of child grammar corpora. We investigated two sets of longitudinal corpora in two unrelated languages (English and Chinese). Both languages display a clear pattern. Younger children from (around) 2 to (around) 4 years of age strongly prefer (in terms of frequency of production in spontaneous speech) 2-merge over 3-merge systems.

Along the same line, we investigated a corpus of utterances of children diagnosed with SLI in English. Though the reduced evidence at our disposal, we detected a trend: 2-merge systems seem to appear earlier with a higher frequency than 3-merge systems.

Future work should investigate in more fine-grained terms the correlation between frequency and complex structures with respect to specific register/genres. Finally, further research should also focus on all N-merge systems in acquisition by enlarging datasets and languages.

Acknowledgements

This research was supported by Science Foundation of Beijing Language and Culture University (supported by "the fundamental Research Funds for the Central Universities") 20YBB06. I am grateful to the audience of the *46th Incontro di Grammatica Generativa* (University of Siena) for useful remarks and questions. I would like to thank Yu Zhao and Emily Stanford for their precious comments on Chinese and English data.

References

Aepli, Noëmi, and Simon Clematide. 2018. "Parsing Approaches for Swiss German." In *Swisstext 2018: 3rd Swiss Text Analytics Conference* (Winterthur, Switzerland, 12-13 June 2018), vol. MMCCXXVI, *CEUR Workshop Proceedings*, ed. by Mark Cieliebak, Don Tuggener, and Fernando Benites, 6-16. <http://ceur-ws.org/Vol-2226/> (06/2021).

Agić, Željko, and Nikola Ljubešić. 2015. "Universal Dependencies for Croatian (that work for Serbian, too)." In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing* (Hissar, Bulgaria, 10-11 September 2015), ed. by Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Hristo Tanev, and Roman Yangarber, 1-8. <https://www.aclweb.org/anthology/W15-5301.pdf> (06/2021).

Ahrenberg, Lars. 2007. "LinES: An English-Swedish Parallel Treebank." In *Proceedings of The 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)* (Tartu, Estonia, 25-26 May 2007), ed. by Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, and Mare Koit, 270-273. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W07-2441> (06/2021).

Badmaeva, Elena, and Francis M. Tyers. 2017. "A Dependency Treebank for Buryat." In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)* (Bloomington, Indiana, 20-21 January 2017), ed. by Markus Dickinson, Jan Hajič, Sandra Kübler, and Adam Przepiórkowski, 1-12. <http://ceur-ws.org/Vol-1779/01badmaeva.pdf> (06/2021).

Batchelor, Colin. 2019. "Universal Dependencies for Scottish Gaelic: Syntax." In *Proceedings of the Celtic Language Technology Workshop 2019* (Dublin, Ireland, 19 August 2019), ed. by Teresa Lynn, Delyth Prys, Colin Batchelor, and Francis Tyers, 7-15. European Association for Machine Translation. <https://www.aclweb.org/anthology/W19-6902.pdf> (06/2021).

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. *Prague dependency treebank 3.0.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics. Prague: Charles University. <http://hdl.handle.net/11858/00-097c-0000-0023-1aaf-3> (06/2021).

Belletti, Adriana. 2004. "Aspects of the Low IP Area." In *The Structure of CP and IP [The Cartography of Syntactic Structures, volume 2]*, ed. by Luigi Rizzi, 16-51.Oxford-New York, NY: Oxford UP.

Belletti, Adriana. 2015. "The Focus Map of Clefts: Extraposition and Predication." In *Beyond Functional Sequence [The Cartography of Syntactic Structures, volume 10]*, ed. by Ur Shlonsky, 42-60. Oxford-New York, NY: Oxford UP.

Belletti, Adriana, and Chris Collins. 2021. *Smuggling in Syntax*. Oxford: Oxford UP.

Belletti, Adriana, Naama Friedmann, Dominique Brunato, and Luigi Rizzi. 2012. "Does Gender Make a Difference? Comparing the Effect of Gender on Children's Comprehension of Relative Clauses in Hebrew and Italian." *Lingua* 122 (10): 1053-1069.

Belletti, Adriana, and Maria T. Guasti. 2015. *The Acquisition of Italian: Morphosyntax and its Interfaces in Different Modes of Acquisition*. Amsterdam-Philadelpia, PA: John Benjamins Publishing Company.

Belletti, Adriana, and Luigi Rizzi. 1981. "The Syntax of 'ne': Some Theoretical Implications." *The Linguistic Review* 1: 117-145.

Berthelot, Frédérique. 2017. *Movement of and out of Subjects in French*. Phd Dissertation, University of Geneva. <https://archive-ouverte.unige.ch/unige:96575> (06/2021).

Bhat, Irshad, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. "Universal Dependency Parsing for Hindi-English Code-Switching." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* v*olume 1 (Long Papers)* (New Orleans, LA, 1-6 June 2018), ed. by Marilyn Walker, Heng Ji, and Amanda Stent, 987-998. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/N18-1090> (06/2021).

Bhat, Riyaz A., Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti M. Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. "The Hindi/Urdu Treebank Project." In *The Handbook of Linguistic Annotation*, ed. by Nancy Ide, and James Pustejovsky, 659-697.Dodrecht-New York, NY: Springer Press.

Bianchi, Valentina, and Mara Frascarelli. 2010. "Is Topic A Root Phenomenon?." *Iberia*: *An International Journal Of Theoretical Linguistics* 2 (1): 43-88.

Bianchi, Valentina, Giuliano Bocci, and Silvio Cruschina. 2015. "Focus Fronting and its Implicatures." In *Romance Languages and Linguistic Theory 2013: Selected papers from "Going Romance" Amsterdam 2013*, ed. by Enoch O. Aboh, Jeannette Schaeffer, and Petra Sleeman, 1-20. Amsterdam: John Benjamins.

Biberauer, Theresa, Ander Holmberg, Ian Roberts, and Michelle Sheehan. 2010. *Parametric Syntax: Null Subjects in Minimalist Theory.* Cambridge-New York, NY: Cambridge UP.

Bielinskienė, Agnė, Loïc Boizou, Jolanta Kovalevskaitė, and Erika Rimkutė. 2016. "Lithuanian Dependency Treebank Alksnis." In *Human Language Technologies – The Baltic Perspective*, ed. by Inguna Skadiņa, and Roberts Rozis, 107-114. Amsterdam: Ios Press. doi: 10.3233/978-1-61499-701-6-107.

Boeckx, Cedric, and Constantina Theofanopoulou. 2014. "A Multidimensional Interdisciplinary Framework for Linguistics: The Lexicon as a Key Study." *Journal of Cognitive Science* 15 (4): 403-420. doi: 10.17791/jcs.2014.15.4.403.

Bonan, Caterina. 2019. *On Clause-Internally Moved Wh-Phrases: Wh-To-Foc, Nominative Clitics, And The Theory Of Northern Italian Wh-In Situ*. PhD Dissertation, University of Geneva. <https://archive-ouverte.unige.ch/unige:119060> (06/2021).

Borer, Hagit. 1994. "The Projection of Arguments". *University of Massachusetts Occasional Papers in Linguistics* 20 (1): 17-49. <http://www.scholarworks.umass.edu/umop/vol20/iss1/3> (06/2021).

Borer, Hagit, and Kenneth Wexler. 1987. "The Maturation of Syntax." In *Parameter Setting. Studies in Theoretical Psycholinguistics*, ed. by Thomas Roeper, and Edwin Williams, 123-172. Dordrecht: Springer. doi: 10.1007/978-94-009-3727-7_6.

Borges Völker, Emanuel, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. "HDT-UD: A Very Large Universal Dependencies Treebank for German." In *Proceedings Of The Third Workshop On Universal Dependencies (Udw, Syntaxfest 2019)* (Paris, France, 26-30 August 2019), ed. by Alexandre Rademaker, and Francis Tyers, 46-57. Cambridge, MA: MIT Press. <https://www.aclweb.Org/anthology/W19-8006> (06/2021).

Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. "Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank." In *Proceedings Of The 7th Linguistic Annotation Workshop And Interoperability With Discourse* (Sofia, Bulgaria, 8-9 August 2013), ed. by Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper, 61-69. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W13-2308/> (06/2021).

Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. "Soft Constraints Mirror Hard Constraints: Voice And Person In English And Lummi." In *Proceedings of the LFG-01 Conference,* ed. by Miriam Butt, and Tracy H. King, 13-32. Stanford, CA: CSLI Publications.

Bross, Fabian. 2020. "Encoding Different Types Of Topics And Foci In German Sign Language. A Cartographic Approach To Sign Language Syntax." In *Glossa: A Journal of General Linguistics* 5 (1): 108. doi: 10.5334/Gjgl.1094.

Caloi, Irene. 2013. "The Comprehension Of Relative Clauses In Patients With Alzheimer's Disease." *Studies in Linguistics* 5: 5-24.

Cardinaletti, Anna. 2004. "Toward A Cartography Of Subject Positions." In *The Structure of CP and IP:The Cartography of Syntactic Structures, Volume 2*, ed. by Luigi Rizzi, 115-165. Oxford, New York, NY: Oxford UP.

Cecchini, Flavio M., Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. "Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies." In *Proceedings Of The Universal Dependencies Workshop 2018 (Udw 2018)* (Brussels, Belgium, 1 November 2018), ed. by Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, 27-36. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W18-6004.pdf> (06/2021).]

Çetinoğlu, Özlem, and Çagrı Çöltekin. 2019. "Challenges of Annotating a Code-Switching Treebank." In *Proceedings Of The 18th International Workshop On Treebanks And Linguistic Theories (TLT, Syntaxfest 2019)* (Paris, France, 28-29 August 2019), ed. by Marie Candito, Kilian Evang, Stephan Oepen, and Djamé Seddah, 82-90. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W19-7809.pdf> (06/2021).

Chesi, Cristiano. 2012. *Competence And Computation: Toward A Processing Friendly Minimalist Grammar*. Padova: Unipress.

Chesi, Cristiano. 2015. "On Directionality Of Phrase Structure Building." *Journal of Psycholinguistic Research* 44 (1): 65-89.

Chesi, Cristiano, and Paolo Canal. 2019. "Person Features and Lexical Restrictions in Italian Clefts." In *Frontiers in Psychology* 10. doi: 10.3389/fpsyg.2019.02105.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1981. *Lectures On Government And Binding: The Pisa Lectures*. Dordrecht: Foris Publications.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2013. "Problems Of Projection." *Lingua* 130: 33-49.

Cignarella, Alessandra T., Cristina Bosco, Viviana Patti, and Mirko Lai. 2018. "Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7-12 May 2018), ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, 4204-4211. Paris : European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1664.pdf> (06/2021).

Cinque, Guglielmo. 2014. *Typological Studies: Word Order and Relative Clauses*. New York, NY: Routledge.

Collins, Chris. 2005. "A Smuggling Approach To Raising In English." *Linguistic Inquiry* 36 (2): 289-298.

Conti-Ramsden, Gina, and Jane J. Dykins. 1991. "Mother-Child Interactions with Language-Impaired Children and their Siblings." *British Journal Of Disorders Of Communication* 26 (3): 337-354.

Conti-Ramsden, Gina, Graeme D. Hutcheson, and John Grove. 1995. "Contingency And Breakdown: Specific Language Impaired Children's Conversations With Their Mothers And Fathers." *Journal Of Speech And Hearing Research* 38 (6): 1290-1302. doi: 10.1044/jshr.3806.1290.

Conti-Ramsden, Gina, and Melanie Jones, 1997. "Verb Use in Specific Language Impairment." *Journal Of Speech And Hearing Research* 40 (6): 1298-1313. doi: 10.1044/jslhr.4006.1298.

Dehaene, Stanislas, Florent Meyniel, Catherine Wacongne, Liping Wang, and Christophe Pallier. 2015. "The Neural Representation Of Sequences: From Transition Probabilities To Algebraic Patterns And Linguistic Trees." *Neuron* 88 (1): 2-19.

Deng, Xiangjun, Ziyin Mai, and Virginia Yip. 2015. *An Aspectual Account Of Ba And Bei Constructions In Child Mandarin.* Paper Presented At The International Symposium On Psycholinguistics Of Second Language Acquisition And Bilingualism, Chinese University Of Hong Kong.

Deng, Xiangjun, and Virginia Yip, 2018. "A Multimedia Corpus Of Child Mandarin: The Tong Corpus." *Journal Of Chinese Linguistics* 46 (1): 69-92.

De Freitas, Marília F.P. 2017. *A Posse Em Apurinã: Descrição De Construções Atributivas E Predicativas Em Comparação Com Outras Línguas Aruák*. PhD Dissertation. Parà: Federal University of Pará.

Di Domenico, Elisa, Ioli Baroncini, and Andrea Capotorti. 2020. "Null and Overt Subject Pronouns in Topic Continuity and Topic Shift: An Investigation of the Narrative Productions of Italian Natives, Greek Natives and Near-native Second Language Speakers of Italian with Greek as a First language." *Glossa: A Journal Of General Linguistics* 5 (1): 117. doi: 10.5334/gjgl.1009.

Droganova, Kira, Olga Lyashevskaya, and Daniel Zeman. 2018. "Data Conversion And Consistency Of Monolingual Corpora: Russian Ud Treebanks." In *Proceedings Of The 17th International Workshop On Treebanks And Linguistic Theories (Tlt 2018)*, (Oslo, Norway, 13-14 December 2018), ed. by Dag Haug, Stephan Oepen, Lilja Øvrelid, Marie Candito, and Jan Hajic, 52-65. Linköping: Linköping University Electronic Press.

Dryer, Matthew S. 2009. "Problems Testing Typological Correlations with the Online WALS." *Linguistic Typology* 13 (1): 121-135. doi: 10.1515/lity.2009.007.

Dryer, Matthew S. 2013. "Expression Of Pronominal Subjects." In *The World Atlas Of Language Structures Online*, ed. by Matthew S. Dryer, and Martin Haspelmath. <http://wals.info/chapter/101> (06/2021).

Dryer, Matthew S., and Martin Haspelmath. 2013. *WALS Online*. Leipzig: Max Planck Institute For Evolutionary Anthropology. <https://wals.info/> (06/2021).

Durrleman, Stephanie, Loyse Hippolyte, Sandrine Zufferey, Katia Iglesias, and Nouchine Hadjikhani. 2015. "Complex Syntax in Autism Spectrum Disorders: A Study of Relative Clauses." *International Journal Of Language and Communication Disorders* 50 (2): 260-267.

Eckhoff, Hanne M., and Aleksandrs Berdicevskis. 2015. "Linguistics vs. Digital Editions: The Tromsø Old Russian and OCS Treebank." *Scripta and E-Scripta* 14-15: 9-25.

Ephrem Seyoum, Binyam, Yusuke Miyao, and Baye Yimam. 2018. "Universal Dependencies For Amharic." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7-12 May 2018), ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, 2216-2222. Paris: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1350> (06/2021).

Frascarelli, Mara. 2007. "Subjects, Topics and the Interpretation of Referential Pro." *Natural Language and Linguistic Theory* 25 (4): 691-734.

Frascarelli, Mara, and Marco Casentini. 2019. "The Interpretation of Null Subjects in a Radical Pro-drop Language: Topic Chains and Discourse-semantic Requirements in Chinese." *Studies In Chinese Linguistics* 40 (1): 1-45. doi: 10.2478/scl-2019-0001.

Frascarelli, Mara, and Roland Hinterhölzl. 2007. "Types of Topics in German and Italian." In *On Information Structure, Meaning and Form: Generalizations Across Languages*, ed. by Kerstin Schwabe, and Susanne Winkler, 87-116. Amsterdam-Philadelphia, PA: John Benjamins Publishing Company.

Frauenfelder, Ulrich, Juan Segui, and Jacques Mehler. 1980. "Monitoring Around the Relative Clause." *Journal of Verbal Learning and Verbal Behavior* 19 (3): 328-337.

Friedmann, Naama, Adriana Belletti, and Luigi Rizzi. 2009. "Relativized Relatives: Types of Intervention in the Acquisition of A-bar Dependencies." *Lingua* 119 (1): 67-88.

Friedmann, Naama, Adriana Belletti, and Luigi Rizzi. 2020. "Growing Trees. The Acquisition of the Left Periphery." Manuscript. University of Tel Aviv, University of Siena, Collège de France. <https://ling.auf.net/lingbuzz/005369> (06/2021).

Gallego, Ángel J. 2008. "Four Reasons to Push Down the External Argument." Manuscript. Universitat Autònoma De Barcelona. <http://filcat.uab.cat/clt/membres/professors/gallego/pdf/external_argument.pdf.> (06/2021).

Gökirmak, Memduh, and Francis M. Tyers. 2017. "A Dependency Treebank for Kurmanji Kurdish." In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)* (Pisa, Italy, 18-20 September 2017), ed. by Simonetta Montemagni, and Joakim Nivre, 64-72. Linköping: Linköping University Electronic Press. <https://www.aclweb.org/anthology/W17-6509.pdf> (06/2021).

Grillo, Nino. 2008. *Generalized Minimality: Sinctactic Underspecification in Broca's Aphasia*. PhD Dissertation. York: University of York.

Guasti, Maria T. 2002. *Language Acquisition – The Growth of Grammar*. Cambridge, MA: MIT Press.

Guasti, Maria T. 2017. *Language Acquisition – The Growth of Grammar*. Cambridge, MA: MIT Press.

Gulordava, Kristina, and Paola 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)*, *Newark, Delaware,* ed. by Irene Vogel, 110-132. Amsterdam-Philadelphia, PA: John Benjamins Publishing Company. doi: 10.1075/rllt.16.08gul.

Haegeman, Liliane. 1990. "Non-Overt Subjects in Diary Contexts." In *Grammar in Progress: Glow Essays for Henk Van Riemsdijk,* ed. by Joan Mascaró, and Marina Nespor, 167-174. Dodricht: Foris Publications.

Hale, Ken, and Samuel J. Keyser. 1998. "The Basic Elements of Argument Structure." In *MIT Working papers in linguistics,* vol. XXXII, *Papers from the UPenn/MIT Roundtable on Argument Structure and Aspect*, ed. by Heidi Harley. 73-118. Cambridge, MA: MIT Press.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The Weka Data Mining Software: An Update." *ACM SGKDD Explorations Newsletter* 11 (1): 10-18. doi: 10.1145/1656274.1656278.

Haug, Dag T., and Marius L. Jøhndal. 2008. "Creating a Parallel Treebank of the Old Indo-European Bible Translations." In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (Marrakech, Morocco, 1 June 2008), ed. by Caroline Sporleder and Kiril Ribarov, 27-34. Cambridge, MA: MIT Press.

Hellwig, Oliver, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. "The Treebank of Vedic Sanskrit." In *Proceedings of the 12th Language Resources and Evaluation Conference* (Marseille, France, 11-16 May 2020), ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 5137-5146. Paris: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/2020.lrec-1.632.pdf> (06/2021).

Heinecke, Johannes, and Francis M. Tyers. 2019. "Development of a Universal Dependencies treebank for Welsh." In *Proceedings of the Celtic Language Technology Workshop* (Dublin, Ireland, 19 August 2019), ed. by Teresa Lynn, Delyth Prys, Colin Batchelor, and Francis Tyers, 21-31. European Association for Machine Translation. <https://www.aclweb.org/anthology/W19-6904.pdf> (06/2021).

Holmberg, Anders, and Ian Roberts. 2013. "The Syntax-Morphology Relation." *Lingua* 130: 111-131. doi: 10.1016/j.lingua.2012.10.006.

Ibbotson, Paul. 2013. "The Scope of Usage-Based Theory." *Frontiers in Psychology* 4. doi:10.3389/fpsyg.2013.00255.

Jaeggli, Osvaldo, and Kenneth J. Safir. 1989. "The Null Subject Parameter and Parametric Theory." In *The Null Subject Parameter*, ed. by Osvaldo Jaeggli, and Kenneth J. Safir, 1-44. Dordrecht-Boston, MA-London: Kluwer Academic Publisher.

Johannsen, Anders, Héctor Martínez Alonso, and Barbara Plank. 2015. "Universal Dependencies For Danish." In *Proceedings of The Fourteenth Internationall Workhop on Treebank and Linguistic Theories (TLT14)* (Warsaw, Poland, 11-12 December 2015), ed. by Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, 157-167. <http://tlt14.ipipan.waw.pl/proceedings/> (06/2021).

Kam, Carla H., and Elissa L. Newport. 2005. "Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change." *Language Learning and Development* 1 (2): 151-195.

Kayne, Richard. 1994. *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Koopman, Hilda, and Dominique Sportiche. 1991. "The Position of Subjects" *Lingua* 85 (2-3): 211-258. doi: 10.1016/0024-3841(91)90022-W. <https://www.sciencedirect.com/science/article/pii/002438419190022W> (06/2021).

Larson, Richard K. 1988. "On the Double Object Construction." *Linguistic Inquiry* 19 (3): 335-391.

Longobardi, Giuseppe. 1994. "Reference and Proper Names: a Theory of N-Movement in Syntax and Logical Form." *Linguistic Inquiry* 25 (4): 609-665.

Luukko, Mikko, Aleksi Sahala, Sam Hardwick, and Krister Lindén. 2020. "Akkadian Treebank for Early Neo-Assyrian Royal Inscriptions." In *Proceedings of the 19th International Workshop on Treebanks*

and Linguistic Theories – Düsseldorf (Düsseldorf, Germany, 27-28 October, 2020), ed. by Kilian Evang, Laura Kallmeyer, Rafael Ehren, Simon Petitjean, Esther Seyffarth, and Djamé Seddah, 124-134. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/2020.tlt-1.11> (06/2021).

MacWhinney, Brian. 2000a. *The CHILDES project: Tools for Analyzing Talk: Volume I: Transcription Format and Programs*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, Brian. 2000b. *The CHILDES Project: Tools for Analyzing Talk: Volume II: The Database*. Mahwa, NJ: Lawrence Erlbaum Associates.

Makazhanov, Aibek, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015. "Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A Report." In *Proceedings of the International Conference "Turkic Language Processing". TurkLang-2015* (Kazan, Russia, 17-19 September 2015), 338-350. Kazan: Tartan Academy of Sciences. <http://www.turklang.net/wp-content/uploads/2017/05/proceedings.pdf > (06/2021).

Martini, Karen, Adriana Belletti, Santi Centorrino, and Maria Garraffa. 2020. "Syntactic Complexity in the Presence of an Intervener: The Case of an Italian Speaker with Anomia." *Aphasiology* 34 (8): 1016-1042. doi: 10.1080/02687038.2019.1686744.

Maslinsky, Kirill. 2014. "Daba: a Model and Tools for Manding Corpora." In *Proceedings Of Talaf 2014: Traitement Automatique Des Langues Africaines* (Marseille, France, 1 July 2014), ed. by Mathieu Mangeot, and Fatiha Sadat, 114-122. Association pour le Traitement Automatique des Langues. <https://www.aclweb.org/anthology/W14-6502> (06/2021).

Merlo, Paola. 2016. "Quantitative Computational Syntax: Some Initial Results." *Italian Journal Of Computational Linguistics* 2 (1): 11-30.

Merlo, Paola, and Sarah 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:10.5334/gjgl.149.

Merlo, Paola, and Suzanne S. Stevenson. 1998. "What Grammars Tell Us About Corpora: The Case of Reduced Relative Clauses". In *Proceedings of the Sixth Workshop on Very Large Corpora* (Montreal, Canada, 15-16 August 1998), ed. by Eugene Charniak, 134-142. Montreal: University of Montreal. <https://www.aclweb.org/anthology/W98-1116> (06/2021).

Moro, Andrea. 1997. "Dynamic Antisymmetry: Movement as a Symmetry-breaking Phenomenon." *Studia Linguistica* 51 (1): 50-76.

Murphy, Elliot. 2015. "The Brain Dynamics of Linguistic Computation." *Frontiers in Psychology* 6. doi:10.3389/fpsyg.2015.01515.

Neeleman, Ad, and Kriszta Szendrői. 2007. "Radical Pro Drop and the Morphology of Pronouns." *Linguistic Inquiry* 38 (4): 671-714.

Nivre, Joakim. 2015. "Towards a Universal Grammar for Natural Language Processing." In *International Conference on Intelligent Text Processing and Computational Linguistics: 16th International Conference, CICLing 2015, Proceedings, Part I ,* (Cairo, Egypt, April 14-20, 2015), ed. by Alexander Gelbukh, 3-16. Cham: Springer.

Ojha, Atul Kr., and Daniel Zeman. 2020. "Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri." In *Proceedings of the WILDRE5 – 5th Workshop on Indian Language Data: Resources and Evaluation* (Marseille, France, 11-16 May 2020), ed. by Girish Nath Jha, Kalika Bali, Sobha L., S.S. Agrawal, and Atul Kr. Ojha, 33-38. Paris: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/2020.wildre-1.7> (06/2021).

Partanen, Niko, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. "The First Komi-Zyrian Universal Dependencies Treebanks." In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (Brusselles, Belgium, 1 November 2018), ed. by Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, 126-132. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W18-6015> (06/2021).

Perlmutter, David M. 1978. "Impersonal Passives and the Unaccusative Hypothesis." In *Proceedings of the 4th Annual Meeting of The Berkeley Linguistics Society* (BLS) 4: 157-189. University of California: Escolarship. <https://escholarship.org/uc/item/73h0s91v> (06/2021).

Pollock, Jean-Yves. 1989. "Verb Movement, Universal Grammar, and the Structure of IP". *Linguistic Inquiry* 20 (3): 365-424.

Preminger, Omer. 2008. "Argument Externality." Manuscript. <https://ling.auf.net/lingbuzz/000569> (06/2021).

Prokopidis, Prokopis, and Haris Papageorgiou. 2017. "Universal Dependencies for Greek." In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)* (Gohenburg, Sweden, 22 May 2017), ed. by Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, 102-106. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W17-0413> (06/2021).

R Development Core Team. 2016. *R: A Language And Environment For Statistical Computing*. Vienna: R Foundation For Statistical Computing.

Ramchand, Gillian. 2008. *Verb Meaning and the Lexicon: A First Phase Syntax*. Cambridge: Cambridge UP.

Rizzi, Luigi. 1982. *Issues In Italian Syntax*. Dordrecht: Foris Publications.

Rizzi, Luigi. 1997. "The Fine Structure Of The Left Periphery." In *Elements Of Grammar: Handbook of Generative Syntax*, ed. by Liliane Haegeman, 281-337. Dordrecht: Kluwer Academic Publishers.

Rizzi, Luigi. 2006. "On the Form of Chains: Criterial Positions and ECP Effects." In *Current Studies in Linguistics*, vol. XLII, *Wh-Movement: Moving On*, ed. by Lisa Cheng, and Norbert Corver, 97-133. Cambridge, MA: MIT Press.

Rizzi, Luigi. 2015a. "Cartography, Criteria, And Labeling." In *Beyond Functional Sequence [The Cartography Of Syntactic Structures, Volume 10]*, ed. by Ur Shlonsky, 314-338. Oxford-New York, NY: Oxford UP. doi: 10.1093/acprof:oso/9780190210588.003.0017.

Rizzi, Luigi 2015b. "Notes on Labeling and Subject Positions." In *Structures, Strategies and Beyond – Studies In Honour Of Adriana Belletti*, ed. by Cornelia Hamann, Elisa Di Domenico, and Simona Matteini, 17-46. Amsterdam: John Benjamins Publishing Company.

Rizzi, Luigi. 2016. "Monkey Morpho-Syntax and Merge-Based Systems." *Theoretical Linguistics* 42 (1-2): 139-145. doi: 10.1515/tl-2016-0006.

Rizzi, Luigi, and Giuliano Bocci. 2017. "Left Periphery of the Clause: Primarily Illustrated for Italian." In *The Blackwell Companion To Syntax*, ed. by Martin Everaert, and Henk Van Riemsdijk, 1-30. Hoboken, NJ: John Wiley and Sons.

Rueter, Jack. 2018. *Erme UD Moksha (Version v1.0)*. doi:10.5281/zenodo.1156112.

Rueter, Jack, Niko Partanen, and Larisa Ponomareva. 2020. "On the Questions in Developing Computational Infrastructure for Komi-Permyak." In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*,(Wien, Austria, 10-11 January 2020), ed. by Tommi A Pirinen, Francis M. Tyers, and Michael Rießler, 15-25. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/2020.iwclul-1.3> (06/2021).

Rueter, Jack, and Francis M. Tyers. 2018. "Towards an Open-source Universal-dependency Treebank for Erzya." In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages* (Helsinki, Finland, 8-9 January 2018), ed. by Tommi A. Pirinen, Michael Rießler, Jack Rueter, Trond Trosterud, and Francis M. Tyers, 106-118. Cambridge, MA: MIT Press <https://www.aclweb.org/anthology/W18-0210> (06/2021).

Samo, Giuseppe. 2019. "Cartography and Locality in German: a Quantitative Study with Dependency Structures." *Rivista Di Grammatica Generativa/Research in Generative Grammar* 41 (5): 1-26. <http://lear.unive.it/jspui/handle/11707/7782> (06/2021).

Samo, Giuseppe, and Paola 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, 26-30 August 2019), ed. by Xinying Chen, and Ramon Ferrer-i-Cancho, 46-56. Cambridge, MA: MIT Press. doi: 10.18653/v1/w19-7906.

Samo, Giuseppe, Yu Zhao, and Gaya Gamhewage. 2020. "Syntactic Complexity of Learning Content in Italian for Covid-19 Frontline Responders: A Study on WHO's Emergency Learning Platform." *Verbum* 11. doi:10.15388/Verb.15.

Sanchez, Alessandro, Stephan C. Meylan, Mika Braginsky, Kyle E. MacDonald, Daniel Yurovsky, and Michael C. Frank. 2019. "Childes-db: A Flexible and Reproducible Interface to the Child Language Data Exchange System." *Behavior Research Methods* 51: 1928-1941.

Schlenker, Philippe, Emmanuel Chemla, Anne M. Schel, James Fuller, Jean-Pierre Gautier, Jeremy Kuhn, Dunja Veselinović, Kate Arnold, Cristiane Cäsar, Sumir Keenan, Alban Lemasson, Karim Ouattara, Robin Ryder, and Klaus Zuberbühler. 2016. "Formal Monkey Linguistics." *Theoretical Linguistics* 42 (1-2): 1-90. doi: 10.1515/tl-2016-0001.

Shim, Ji Y. 2016. "Mixed Verbs in Code-Switching: The Syntax of Light Verbs." *Languages* 1 (1): 8. doi: 10.3390/languages1010008.

Si, Fuzhen. 2019. "A Cartographic Study of Light Verb Constructions." *Yuwen Xuexi (Literacy Study)*1: 1-20.

Stanford, Emily N. 2020. *The Language-Cognition Interface: Executive Functions and Syntax in Atypical Development.* PhD Dissertation, University of Geneva. doi: 10.13097/archive-ouverte/unige:144700.

Prévost, Sophie, and Achim Stein. 2013. "Syntactic Annotation of Medieval Texts: the Syntactic Reference Corpus of Medieval French (SRCMF)." In *New Methods in Historical Corpora: Corpus Linguistics and International Perspectives on Language*, ed. by Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, 275-282. Tübingen: Gunter Narr Verlag.

Tsarfaty, Reut. 2013. "A Unified Morpho-Syntactic Scheme of Stanford Dependencies." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Sofia, Bulgaria, 4-9 August 2013), ed. by Hinrich Schuetze, Pascale Fung, and Massimo Poesio, 578-584. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/P13-2103> (06/2021).

Tyers, Francis M., and Vinit Ravishankar. 2018. "A Prototype Dependency Treebank for Breton." In *Actes de la Conférence TALN* (Rennes, France, 14-18 May 2018)*, vol 1 − Articles longs, articles courts de TALN*, ed. by Pascale Sébillot, and Vincent Claveau, 197-204. Rennes: ATALA. <https://www.aclweb.org/anthology/2018.jeptalnrecital-court.1> (06/2021).

Tyers, Francis M., and Mariya Sheyanova. 2017. "Annotation Schemes in North Sámi Dependency Parsing." In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages* (St. Petersburg, Russia, 23-24 January 2017), ed. by Francis M. Tyers, Michael Rießler, Tommi A. Pirinen, and Trond Trosterud, 66-75. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W17-0607 > (06/2021).

Tyers, Francis M., Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. "Multi-source Synthetic Treebank Creation for Improved Cross-lingual Dependency Parsing." In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (Brussels, Belgium, 1 November 2018), ed. By Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, 144-150. Cambridge, MA: MIT Press. <https://www.aclweb.org/anthology/W18-6017> (06/2021).

Türk, Utku, Furkan Atmaca, Şaziye B. Özateş, Gözde Berk, Seyyit T. Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2020. *Resources For Turkish Dependency Parsing: Introducing The Boun Treebank And The Boat Annotation Tool*. Manuscript. <https://arxiv.org/abs/2002.10416 (06/2021).

Velldal, Erik, Lilja Øvrelid, and Petter Hohle. 2017. "Joint UD Parsing of Norwegian Bokmål and Nynorsk." In *Proceedings of the 21st Nordic Conference on Computational Linguistics* (Gothenburg, Sweden, 22-24 May 2017), ed. by Jörg Tiedemann, 1-10. Linköping: Linköping University Electronic Press. <https://www.aclweb.org/anthology/W17-0201.pdf > (06/2021).

Volk, Martin, Johannes Graën, and Elena Callegaro. 2014. "Innovations in Parallel Corpus Search Tools." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (Reykjavik, Iceland, 26-31 May 2014), ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1-7. Paris: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2014/pdf/504_Paper.pdf.> (06/2021).

Wells, Gordon. 1981. *Learning Through Interaction: The Study of Language Development*. Cambridge: Cambridge UP. doi: 10.1017/CBO9780511620737.

Yang, Charles. 2013. "Ontogeny and Phylogeny of Language." *Proceedings of the National Academy of Sciences* 110 (16): 6324-6327.

Yang, Charles. 2015. *Generalization and Probability Matching*. Manuscript. University of Pennsylvania. <https://www.ling.upenn.edu/~ycharles/probmatch.pdf> (06/2021).

Yavrumyan, Marat M. 2019. "Universal Dependencies for Armenian." (Paris, France, 3-5 October 2019). In *International Conference on Digital Armenian*. Paris: Institut National des Langues et Civilisations Orientales (INALCO).

Zeman, Daniel. 2017. "Slovak Dependency Treebank in Universal Dependencies." *Jazykovedný Časopis* 68 (2): 385-395.

Zeman, Daniel, Joakim Nivre, and Mitchell M. Abrams, *et al.* 2020. *Universal Dependencies 2.7*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics. Prague: Charles University. <http://hdl.handle.net/11234/1-3424> (06/2021).

Zhang, Li, and Jing Zhou. 2009. "The Development of Mean Length of Utterance in Mandarin-speaking Children (汉语儿童平均语句长度发展研究)." In *The Application and Development of International Corpus-based Research Methods* (汉语儿童语言发展研究: 国际儿童语料库研究方法的应用与发展), ed. by Jing Zhou, 40-58. Beijing: Education Science Publishing House (教育科学出版社).

Zhao, Yu, Giuseppe Samo, Heini Utunen, Oliver Stucke, and Gaya Gamhewage. 2021. "Evaluating Complexity of Digital Learning in a Multilingual Context: A Cross-Linguistic Study on WHO's Emergency Learning Platform". In *Public Health and Informatics*, vol. CCLXXXI of *Studies of Health Technologies and Informatics* ed. by John Mantas, Lăcrămioara Stoicu-Tivadar, Catherine Chronaki, Arie Hasman, Patrick Weber, Parisis Gallos, Mihaela Crișan-Vida, Emmanouil Zoulias, and Oana S. Chirila, 516-517. Amsterdam: IOS Press. doi: 10.3233/SHTI210222.