Word order variation in Mbyá Guaraní

Angelika Kiss
Department of Linguistics
University of Toronto
angelika.kiss@mail.utoronto.ca

Guillaume Thomas
Department of Linguistics
University of Toronto
guillaume.thomas@utoronto.ca

Abstract

This paper presents the preliminary results of a multifactorial analysis of word order in Mbyá Guaraní, a Tupí-Guaraní language spoken in Argentina, Brazil and Paraguay, based on a corpus of written narratives with multiple layers of annotation. Our goals are to assess the validity of previous claims about Mbyá word order (Martins, 2003; Dooley, 1982; Dooley, 2015), and to explore the effects of different types of factors on the position of core arguments relative to their verb. We show that SV and VO are the most frequently attested orders in matrix clauses and that subordinate clauses favour the OV order. Grammatical function, givenness and clause type (root vs subordinate) are found to be significant predictors of core argument position. We identify differences in object position between Mbyá and Paraguayan Guaraní (Tonhauser and Colijn, 2010), and we argue that these differences support Dietrich (2009)'s proposal that Tupí-Guaraní languages are undergoing a change in word order from OV to VO, induced by contact with Spanish and Portuguese.

1 Introduction

This paper presents the preliminary results of a multifactorial analysis of the relative order of subject, object and verb in Mbyá Guaraní, a Tupí-Guaraní language spoken in Argentina, Brazil and Paraguay, which is closely related to Paraguayan Guaraní. To the best of our knowledge, Mbyá word order has only been investigated by Martins (2003), and by Dooley (1982, 2008, 2015). However, these studies do not include detailed reports of word order frequencies, nor do they engage in quantitative modelling of word order variation.

A first goal of the study is to provide statistics that will put the description of word order in Mbyá on a more solid foundation. A second goal is to explore constraints on word order variation in the language through multifactorial techniques. More precisely, we ask what factors affect the position of core arguments relative to their verb, and whether these factors are predominantly syntactic (clause type, grammatical function), discourse-pragmatic (givenness), lexical (animacy, transitivity) or related to processing (argument length). To this end, we annotated a corpus of 1,046 sentences with interlinear glosses, parts of speech tags, syntactic dependency relations and coreference relations, which forms the basis of the present study.

We compare our results to the findings of Tonhauser and Colijn (2010), who investigated subject and object placement in Paraguayan Guaraní. We find notable differences between these two languages, which we interpret in the light of Dietrich (2009)'s analysis of word order change in the Tupí-Guaraní family.

2 Some relevant aspects of Mbyá grammar

Mbyá is a head-marking language. There is no case marking on nouns. Verbs agree in person and number with their core arguments. Intransitive verbs belong to one of two classes, called active and inactive, which use different paradigms of prefixes to cross-reference their subject, as illustrated by the following examples:[1]

(1) a. Xee a- a ju ma.
I A1.SG- go again already
'*I am already going again.*' (Dooley 2015)

b. Xe- kangy vaipa.
B1.SG- feel_weak very
'*I feel very weak.*' (Dooley 2015)

With transitive verbs, the active paradigm is used to cross-reference subjects, and the inactive paradigm is used to cross-reference objects. However, only one argument can be cross-referenced.[2] If both arguments are third person, the subject is cross-referenced. Otherwise, the highest argument on the person hierarchy 1 > 2 > 3 is cross-referenced. In the following example, the verb *xe-r-exa* cross-references its 1^{st} person object. Its implicit subject must be 2^{nd} or 3^{rd} person:

(2) Xe- r- exa.
B1.SG- R- see
'*They/(s)he/you saw me.*' (Dooley, 2015)

Note that Mbyá is a pro-drop language. All core arguments can be omitted, even if they are not cross-referenced on the verb, as illustrated in example (2) for the subject.

Dooley (1982) reports that SVO is the unmarked order, and that SOV, OSV and OVS orders are also attested. Martins (2003) argues that both SOV and SVO are basic word orders, the latter being more prevalent among younger speakers. However, Martins reports that all six permutations of the subject, verb and object were accepted by native speakers.

(3) kuee Maria o- jogua jety (SVO)
yesterday Maria A3 buy potato
'Yesterday Maria bought potatoes'

a. kuee Maria jety o-jogua (SOV)

b. kuee jety o-jogua Maria (OVS)

c. kuee jety Maria o-jogua (OSV)

d. kuee o-jogua jety Maria (VOS)

e. kuee o-jogua Maria jety (VSO) (Martins 2003, p. 154)

Note that Dooley (1982)'s observations are based on his description of Mbyá in the Rio das Cobras community in the Brazilian state of Paraná, while Martins (2003) describes the language spoken in the Morro dos Cavalos and Maciambu communities in the state of Santa Catarina, also in Brazil.

3 Corpus Construction

The corpus used in the present study consists of narratives written between 1976 and 1990 by two Mbyá speakers from the Rio das Cobras community in Paraná, Brazil. These narratives were collected and interlinearized by Robert Dooley. This corpus is available on the Archive of the Indigenous Languages of the America (Dooley, 2011).

[1] Glosses: A1.SG: first person singular 'active' inflection; B1: first person singular 'inactive' inflection; R: linking morpheme.

[2] With the exception of combinations of 1^{st} person subject and 2^{nd} person object, which are cross-referenced with a portmanteau prefix *ro-*.

The 33 narratives used in this study contain 1046 sentences and 11771 tokens. One author, Nelson Florentino, contributed more than 95% of the tokens. The other narratives were written by Darci Pires de Lima.

The corpus was annotated by the authors and research assistants[3]. It contains five layers of annotation: interlinear morphological glosses, parts of speech tags, syntactic dependency relations, coreference annotation and animacy annotation.

Dooley's interlinearization was revised in SIL FieldWorks Language Explorer (Black and Simons, 2008). The interlinearization includes morphological segmentation and glosses, syntactic category annotation using language specific tags, and a free translation into Brazilian Portuguese.

Syntactic annotation was done by the authors in dependency grammar, in the Universal Dependency v2.4 framework (Nivre et al., 2019). Universal POS tags and morphological features were converted automatically from the language specific POS tags and glosses included in the interlinearization layers. Dependency relations were added manually in Arborator (Gerdes, 2013). While the syntactic annotation of Mbyá in Universal Dependencies v2.4 involves a number of non-trivial analytical decisions, the present study only exploits part of the information encoded in the dependency annotation, namely syntactic relations between predicates and their subject and objects, as well as relations of clausal subordination (relative, adverbial and complement clauses). The identification of these relations using Universal Dependency guidelines did not present any particular challenge, and we refer the reader to these guidelines for further information (UD Guidelines, n.d.).

The layer of coreference annotation was created in WebAnno 3 (de Castilho et al., 2016), following Komen (2009)'s annotation guidelines. We understand coreference in a general sense to be a relation between expressions that introduce discourse referents, both referential expressions properly speaking and quantifiers. When a referring expression or a quantifier is used, we call it a *mention* of its discourse referent. Sequences of mentions that have identical or related discourse referents form *referential chains*. Following Bentivoglio (1983), we include implicit mentions of arguments in our referential chains. When an argument is dropped or only expressed in the form of a cross-reference marker on the verb, we consider it an implicit mention and include it in a referential chain. Implicit arguments were annotated by adding null subject and/or object tags on their verb.

A version of the annotated corpus that includes UD dependency annotations is available in a delexicalized form as a part of Universal Dependencies 2.4 (Thomas, 2019). The coreference annotation layer is not yet publicly available at the date of writing of this paper.

4 Data Extraction and Coding Decisions

We exported our corpus to a WebAnno tab-separated file, from which we extracted relevant observations using a Python script. Statistical analysis was performed in R (R Core Team, 2013). Two R data frames were created. In the first one, each observation corresponds to a verb, which is coded for its Transitivity, and for the Word Order of its clause: V (no overt argument), VS, SV, VO, OV, SVO, SOV, OSV, OVS, VSO. VOS order is unattested in the corpus. This data frame includes clausal arguments.

In the second data frame, each observation corresponds to an overt subject or object, which is coded for its position relative to the verb: pre-verbal (XV) or post-verbal (VX). This data frame excludes clausal arguments. In addition, subjects and objects were coded for several independent variables that have been used in quantitative studies of word order (Prince, 1981; Givón, 1983; Ariel, 1988; Hawkins, 1994; Tonhauser and Colijn, 2010; Heylen, 2005): Animacy (animate/inanimate), Clause Type (root/subordinate), Givenness (new/given), Grammatical Function (subject/object), Length (numeric) and Transitivity of the verb (intransitive/transitive).

We excluded dependent verbs in serial verb constructions, as well as identificational constructions and interrogative clauses. Some coding decisions should be noted:

- *Clause Type*: we coded independent clauses and main clauses of direct reported speech as 'root'. Clausal complements, adverbial clauses and relative clauses were all coded as 'subordinate.'

[3] Gregory Antono, Laurestine Bradford, Vidhyia Elango, Jean-François Juneau, Barbara Peixoto, Darragh Winkelman.

- *Givenness*: mentions that do not have an antecedent in the coreference annotation of our corpus were coded as 'new'. We coded as 'given' all mentions that are related to an antecedent through coreference, bridging anaphora or through a partitive relation.
- *Length*: length was coded as the number of characters making up the relevant mention. Since the orthography used in our corpus makes restricted use of digraphs for simple segments, and the phonology of Mbyá does not contrast long and short vowels, this is a reasonable approximation of the number of phonological segments. In several studies, length is coded as number of words of the mention (Jacennik and Dryer, 1992; Siewierska, 1993; Arnold et al., 2000; Rosenbach, 2005), or number of syllables (Heylen, 2005). There have been proposals for substituting length by different measures of syntactic complexity (e.g. the number of syntactic nodes), but length has been argued to be a good enough predictor of syntactic complexity, at least in English (Wasow, 1997; Szmrecsányi, 2004).

5 Analysis

Table 1 presents counts and proportions of word orders in our data set:

		Clause Type		Transitivity	
		root	sub	vi	vt
Word Order	V	591 (62.8)	350 (37.2)	611 (64.9)	330 (35.1)
	SV	304 (76.0)	96 (24.0)	327 (81.8)	73 (18.2)
	VS	61 (89.7)	7 (10.3)	62 (91.2)	6 (8.8)
	OV	186 (79.8)	47 (20.2)		233 (100.0)
	VO	163 (91.6)	15 (8.4)		178 (100.0)
	SOV	33 (80.5)	8 (19.5)		41 (100.0)
	SVO	121 (98.4)	2 (1.6)		123 (100.0)
	OSV	3 (75.0)	1 (25.0)		4 (100.0)
	OVS	7 (87.5)	1 (12.5)		8 (100.0)
	VSO	6 (100)	0 (0.0)		6 (100.0)
Total		1475	527	1000	1002

Table 1: Word Order Overview (including clausal-arguments)

Out of 2002 clauses, 941 have no overt subject or object. Subjects are omitted on 68% of verbs, and objects on 41% of transitive verbs. Only 182 clauses had both overt subjects and objects, out of 1002 transitive clauses.

Let us now restrict our attention to non-clausal arguments. Table 2 gives an overview of our predictors in the subset of clauses with at least one overt non-clausal argument, which includes a total of 944 core arguments. The last column reports the p-value of Chi-Square tests for categorical predictors, and of Kruskal-Wallis tests for numeric predictors (Length). Subjects generally precede their verb, while the distribution of objects is more balanced. Animate and given arguments also tend to occur in pre-verbal position. Post-verbal arguments tend to be longer than pre-verbal ones.

Table 3 presents our predictors separately for subject and object positions. We see that animacy and clause type are not significant predictors of subject position, and only clause type is a significant predictor of object position.

In order to explore the combined effects of our predictors on word order, we turn to multifactorial classification models. We fitted conditional inference tree and random forest models to our data set, using the `ctree` function from the `party` package in R (Hothorn, 2019). These models have the advantage of being appropriate for unbalanced designs with multicollinearity (Tagliamonte and Baayen, 2012). We first fit a conditional inference tree to the whole data set, which lets us explore interactions between our predictors. The tree represented in figure 1 includes all splits that are significant at the level of 0.05.

Position		XV (pre-verbal)	VX (post-verbal)	p
Animacy	animate	578 (82.7)	121 (17.3)	<0.001
	inanimate	143 (58.4)	102 (41.6)	
Clause Type	root	568 (73.9)	201 (26.1)	<0.001
	sub	153 (87.4)	22 (12.6)	
Givenness	given	598 (81.8)	133 (18.2)	<0.001
	new	123 (57.7)	90 (42.3)	
Grammatical Function	S	568 (88.1)	77 (11.9)	<0.001
	O	153 (51.2)	146 (48.8)	
Length	Mean (SD)	7.7 (4.1)	9.4 (4.1)	<0.001
Transitivity	vi	327 (85.2)	57 (14.8)	<0.001
	vt	394 (70.4)	166 (29.6)	

Table 2: Predictors of Argument Position (non-clausal)

		Subjects			Objects		
		XV	VX	p	XV	VX	p
Animacy	animate	533 (88.8)	67 (11.2)	0.027	45 (45.5)	54 (54.5)	0.164
	inanimate	35 (77.8)	10 (22.2)		108 (54.0)	92 (46.0)	
Clause Type	root	461 (86.8)	70 (13.2)	0.035	107 (45.0)	131 (55.0)	<0.001
	sub	107 (93.9)	7 (6.1)		46 (75.4)	15 (24.6)	
Givenness	given	510 (91.1)	50 (8.9)	<0.001	88 (51.5)	83 (48.5)	0.907
	new	58 (68.2)	27 (31.8)		65 (50.8)	63 (49.2)	
Length	Mean (SD)	7.2 (3.7)	9.1 (4.0)	<0.001	9.4 (4.9)	9.5 (4.1)	0.412
Transitivity	vi	327 (85.2)	57 (14.8)	0.006			
	vt	241 (92.3)	20 (7.7)		153 (51.2)	146 (48.8)	

Table 3: Predictors of Argument Position by Grammatical Function (non-clausal)

Examination of the conditional inference tree shows that grammatical function is the most important predictor of core argument placement. We also observe an interaction between grammatical function and givenness. While subjects tend to be preverbal, new subjects are more likely to be post-verbal than given subjects. Grammatical function also interacts with clause type, objects being more likely to be pre-verbal in subordinate than in root clauses.

In order to obtain a more robust assessment of the importance of each variable in predicting word order, we fit a random forest model of 300 trees to our data set, with three variables available for splitting at each node (mtry = 3). Each tree in the forest is built on a random sample of the data set, which serves as a learning-sample for this tree. Some observations, the out-of-bag observations, are held off and used as a built-in test sample for the tree. The prediction accuracy of each tree is calculated on its associated out-of-bag sample (Strobl et al., 2009). The model has an out-of-bag accuracy of 77.9%. Table 4 shows a confusion matrix for the model.

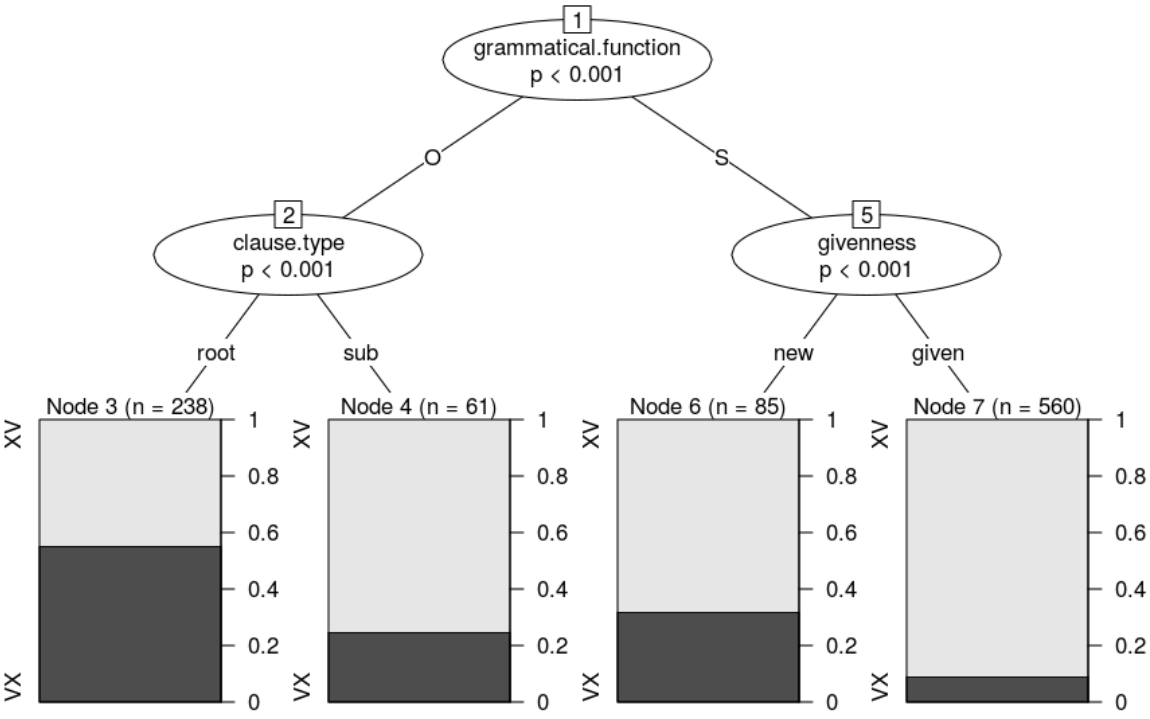

Figure 1: Conditional Inference Tree model of Argument Position

	Predicted: XV	Predicted: VX
Observed: XV	655	66
Observed: VX	142	81

Table 4: Observed values and predictions of the random forest.

Table 5 shows the conditional variable importance (Strobl et al., 2008) for all predictors in our random forest. We see that grammatical function is by far the most important predictor, followed by clause type and givenness. The least important predictors are animacy, length and transitivity. These results are consistent with the conditional inference tree presented in figure 1.

Transitivity	Animacy	Length	Givenness	Clause Type	Grammatical Function
0.00005	0.00209	0.00724	0.00847	0.01053	0.03203

Table 5: Variable Importance in the Random Forest.

The conclusions drawn from the recursive partitioning models are supported by a logistic regression model, which we report in table 6. Again, we observe that grammatical function is the most important predictor, followed by clause type, givenness and transitivity. Animacy and length are not significant predictors in that model.

6 Discussion

We found that while subjects are mostly preverbal in Mbyá (88.1% of all non-clausal subjects in the corpus), the position of objects is more variable, with 51.2% of pre-verbal non-clausal objects and 48.8% of post-verbal non-clausal objects. Given arguments are more likely to be pre-verbal, in keeping with

	Intercept	Length	Animacy (inanimate)	Transitivity (transitive)	Givenness (object)	Cl. Type (sub.)	Gram. Funct. (new)
Coef.	-1.9320	0.0285	-0.1483	-0.6761	0.5916	-1.1623	2.3577
S.E.	0.2314	0.0205	0.2268	0.2796	0.1971	0.2636	0.3104
Z	-8.35	1.39	-0.65	-2.42	3.00	-4.41	7.60
p	<0.0001	0.1634	0.5132	0.0156	0.0027	<0.0001	<0.0001

Table 6: Logistic Regression model of Argument Placement (reference level: pre-verbal).

proposals that old information tend to precede new information across languages (Clark and Clark, 1977; Siewierska, 1993). Objects are more likely to be pre-verbal in subordinate than in root clauses.

Our results support Martins (2003)'s observation that both (S)OV and (S)VO orders are frequently attested in Mbyá, when both arguments are expressed. At the same time, we also found support for Dooley (2015)'s claim that the (S)OV order is more frequent in subordinate clauses.

It is interesting to compare constraints on word order in Mbyá with those that Tonhauser and Colijn (2010) observed for Paraguayan Guaraní. Note that Tonhauser and Colijn (2010) only investigated word order in matrix clauses. While 86.8% of non-clausal subjects are pre-verbal in matrix clauses in our corpus, Tonhauser and Colijn (2010) found that matrix subjects exhibit a greater variability in Paraguayan Guaraní, with only 55% of subjects occurring in pre-verbal position. By contrast, the distribution of objects was found to be less variable in Paraguayan Guaraní, with 95% of direct objects occurring post-verbally compared to Mbyá matrix clauses where 45% of the non-clausal objects are preverbal.

The differences we observed between Mbyá and Paraguayan Guaraní object placement support Dietrich (2009)'s analysis of word order change in Tupí-Guaraní languages. Dietrich argues that Tupí-Guaraní languages are undergoing a change from OV to VO order due in part to contact with Spanish and Portuguese. Of all Tupí-Guaraní languages, Paraguayan Guaraní has had the most sustained contact with Spanish and Portuguese (Melia, 2003), and is also argued to be the language with the most prevalent VO order. Because Mbyá has undergone less contact with Spanish or Portuguese, we expect that OV order will be more frequent overall. Dietrich's hypothesis is also supported by the greater frequency of OV order in subordinate clauses in Mbyá. Since subordinate clauses tend to be more conservative than root clauses (Givón, 1979; Hock, 1986; Bybee, 2002), the lesser frequency of VO order in this environment supports the view that this feature is an innovation in the language.

7 Conclusion

Our study confirmed previous descriptions of word order variation in Mbyá (Martins, 2003; Dooley, 1982; Dooley, 2015). It was found that the position of core arguments relative to the verb is affected by a combination of factors, which are syntactic (clause type, grammatical function) and discourse-pragmatic (givenness). The different frequencies of OV order in Mbyá and Paraguayan Guaraní might be explained by an ongoing change from OV to VO in Tupí-Guaraní languages due to contact with Spanish and Portuguese, which has been more intense in the case of Paraguayan Guaraní.

References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity. *Language*, 76(1):28–55.

Paola Bentivoglio. 1983. Topic Continuity in Spoken Latin-American Spanish. In Talmy Givón, editor, *Topic Continuity in Discourse*, pages 255–312. John Benjamins, Amsterdam/Philadelphia.

Andrew Black and Gary Simons. 2008. The SIL Fieldworks Language Explorer Approach to Morphological Parsing. In *Computational Linguistics for Less Studied Languages: Texas Linguistics Society, 10*, pages 37–55. CSLI Publications.

Joan Bybee. 2002. Main clauses are innovative, subordinate clauses are conservative. Consequences for the nature of constructions Joan L. Bybee and Michael Noonan, editors. *Complex sentences in grammar and discourse: essays in honor of Sandra A. Thompson*, pages 255–312. John Benjamins, Amsterdam/Philadelphia.

H. H. Clark and E. V. Clark. 1977. *Psychology and Language: An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich, New York.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities LT4DH*, pages 11–17, Osaka, Japan. https://webanno.github.io/webanno/.

Wolf Dietrich. 2009. Cambio del orden de palabras en lenguas tupí-guaraníes [Word order change in Tupi-Guarani languages]. *Cadernos de Etnolingüística*, 1:1–11.

Robert A. Dooley. 1982. Options in the pragmatic structuring of Guaraní sentences. *Language*, 58:307–31.

Robert A. Dooley. 2008. Pronouns and topicalization in Guarani texts. Associação Internacional de Lingüística - SIL Brasil, Cuiabá MT.

Robert A. Dooley. 2015. Léxico guarani, dialeto mbyá. Summer Institute of Linguistics.

Robert A. Dooley. 2011 Mbyá Guaraní collection of Robert Dooley. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: text. Access: 100% restricted. PID ailla:119734.

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97.

Talmy Givón. 1979. *On understanding grammar*. New York: Academic Press.

Talmy Givón (editor). 1983. *Topic Continuity in Discourse. A quantitative cross-language study*. John Benjamins, Amsterdam/Philadelphia.

Talmy Givón. 1988. The pragmatics of word-order: predictability, importance and attention. In Michael Hammond, Edith A. Moravcsik, and Jessica R. Wirth, editors, *Studies in Syntactic Typology*, pages 243–285. John Benjamins, Amsterdam/Philadelphia.

John Hawkins. 1994. *A performance theory of order and consituency*. Cambridge University Press, Cambridge, Massachusetts.

Kris Heylen. 2005. A quantitative corpus study of German word order variation. In St. Kepser and M. Reis, editors, *Linguistic evidence: Empirical, theoretical and computational perspectives*, pages 241–264. Mouton de Gruyter, Berlin & New York.

Hans H. Hock. 1986. *Principles of historical linguistics*. Berlin/New York: Mouton deGruyter.

Torsten Hothorn, Kurt Hornik, Carolin Strobl and Achim Zeileis. 2019. *party: A laboratory for recursive part(y)itioning (R package version 1.3–3)*. https://cran.r-project.org/web/packages/party/

Barbara Jacennik and Matthew S. Dryer. 1992. Verb-subject order in Polish. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 209–242. John Benjamins, Amsterdam/Philadelphia.

Erwin R. Komen. 2009. Coreference Annotation Guidelines. http://repository.ubn.ru.nl/bitstream/handle/2066 /78810/78810.pdf.

Marci Fileti Martins. 2003. *Descrição e análise de aspectos de gramática do guarani mbyá [Description and analysis of some grammatical aspects of Guaraní Mbyá]*. Ph.D. thesis, State University of Campinas.

Bartomeu Meliá 2003. *La Lengua Guaraní del Paraguay*. Asunción: CEPAG.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Joakim Nivre, Mitchell Abrams, and Agić Željko et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ellen F. Prince. 1981. Toward a taxonomy of given–new information. In Cole, Peter, editors, *Radical Pragmatics*, pages 223–255. New York: Academic Press.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/.

Anette Rosenbach. 2005. Animacy Versus Weight as Determinants of Grammatical Variation in English. *Language*, 81:613–644.

Anna Siewierska. 1993. Syntactic weight vs information structure and word order variation in Polish. *Journal of Linguistics*, 29:233–265.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9: 307 https://doi.org/10.1186/1471-2105-9-307

Carolin Strobl, James Malley and Gerhard Tutz. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, 14(4):323–348.

Benedikt Szmrecsányi. 2004. On Operationalizing Syntactic Complexity. In G. Purnelle, C. Fairon, and A. Dister, editors, *Le poids des mots. 7es Journées internationales d'Analyse statistique des Données Textuelles*, pages 1031–1039, Louvain-la-Neuve. Presses universitaires de Louvain.

Sali A. Tagliamonte and R. Harald Baayen 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice *Language Variation and Change*, 24:135–178.

Guillaume Thomas. 2019. UD Mbya_Guarani_Dooley, Mbyá Guaraní treebank based on narratives collected by Robert Dooley. In Nivre et al. 2019.

Judith Tonhauser and Erika Colijn. 2010. Word order in Paraguayan Guaraní. *International Journal of American Linguistics*, 76:255–288.

Universal Dependencies n.d. Universal Dependencies Guidelines.
https://universaldependencies.org/guidelines.html

Thomas Wasow. 1997. Remarks on grammatical weight. *Language Variation and Change*, 9:81–105.