Otto-Friedrich-Universität Bamberg
Faculty of Humanities
Department of General Linguistics
Master thesis, 30 ECTS
Examiner: Prof. Dr. Geoffrey Haig
Second examiner: Nils Norman Schiborr
Summer Semester 2022

Historical contingency and typological tendencies in languages of Western Asia: A quantitative study of word order of non-subject constituents

Alexandru Craevschi

Matriculation number: 2045860
MA General linguistics (4)
Hartmannstraße 8, 96050, Bamberg
Tel.: +49 172 4476279
E-mail: alexandru.craevschi@stud.uni-bamberg.de

Otto-Friedrich-Universität Bamberg
Fakultät Geistes- und Kulturwissenschaften
Lehrstuhl für Allgemeine Sprachwissenschaft
Masterarbeit, 30 ECTS
Prüfer: Prof. Dr. Geoffrey Haig
Zweitprüfer: Nils Norman Schiborr
Sommersemester 2022

Historische Kontingenz und typologische Tendenzen in Sprachen Westasiens: Eine quantitative Untersuchung zur Wortstellung von Nicht-Subjekt-Konstituenten

Alexandru Craevschi

Matrikelnummer: 2045860
MA Allgemeine Sprachwissenschaft (4)
Hartmannstraße 8, 96050, Bamberg
Tel.: +49 172 4476279
E-mail: alexandru.craevschi@stud.uni-bamberg.de

Contents

List of Figures

List of Tables

1 Introduction

1.1 Methodological problems and advances in typology

Linguistic typology is arguably one of the most fruitful sub-disciplines of language sciences. It aims at classifying a set of languages according to their structural properties and afterwards explaining the way the data and the observed correlations between some of the grammatical features came into being. These explanations vary greatly from one researcher to the other.

Since the breakthrough publication by Joseph Greenberg (1963), the field of linguistic typology has been concerned with the sampling of languages when designing a cross-linguistic study. Since typologists are most often interested in structural features that are cross-linguistically universal, researchers assemble a set of features from languages that belong to different linguistic families but they also pick languages from different regions of the world. The reason for exclusion based on these principles lies in the processes that drive linguistic change, variation and diversification. That means that closely related languages, be it genetically related or through contact, cannot be considered independent data points. More generally, the same problem in social sciences has the name "Galton's problem" (Naroll, 1961). For instance, Collins (2018, p. 53) mentions that in order to establish causal relations in cross-linguistic research, one has to control for genealogical history of a language family/families to establish reliably whether it was lineage-specific development that accounts for observed features or it is indeed a more universal principle that is independent of specific language family development. As an

example of a good study in this sense, Collins (2018) mentions work by Dunn et al. (2011). However, the paper by Dunn et al. (2011) was criticized by linguists and is rarely taken as a decisive evidence in the matter, as it fails to account for language contact:

> But a more important drawback is that there is no control for language contact. What could be happening is that some Indo-European languages in India have different word orders because of the languages that they are near, such as Dravidian languages, which also have object-verb order and postpositions. (Collins, 2018, p. 53)

An important nuance is that Dunn et al. (2011) tried to measure how languages change states between different features and more importantly, how correlated are the changes between verb and object positions on one side, and adposition and noun phrase on the other side. As authors of the work used phylogenetic methods imported directly from evolutionary biology, there is no way to account for contact in biological phylogenetic trees. Because of that, phylogenetic comparative methods are suitable for certain kinds of linguistic questions but should also be used and interpreted with caution. Up until today, there is no reliable method to account for contact in a language tree, although a new work by Neureiter et al. (2022) looks as an attractive solution for this purpose. The method by Neureiter et al. (2022) estimates borrowings between languages via a separate parameter included in the tree. It does so by taking into account the geographical position of languages. The ultimate product of the method is still a tree, however it additionally contains horizontal edges that reflect borrowing events. Nonetheless, at the moment of writing this paper it is still unclear how to apply the new method to study things like correlated evolution of features, as it was done in Dunn

et al. (2011). It also remains unclear whether already developed phylogenetic comparative methods are compatible with the presence of contact in the trees or require further adaptation for new, more nuanced linguistic trees.

Later in the introduction I will make a short overview of how linguists deal with contact and phylogenetic biases in language development methodologically. This will be important later for understanding the methods used in this study.

Another important aspect of classical typology[1] is the discretization of values of typological features. For comparability purposes and often out of convenience, some structural properties of languages in a sample are discretized, as for example in Dryer (2013c). That means that in order to determine a value of a feature for a particular language, we apply certain criteria for categorization and that allows to have a categorical variable with two or more levels for every language in the sample. Needless to say, this leads to loss of information. The new approaches nowadays extract typological information from corpora, most often from Universal Dependencies (UD) (Zeman et al., 2019) and can, for example, quantify what proportion of clauses in a language are OV or VO, instead of deciding on a single value. Universal Dependencies, along with some other corpora, has paved the way for corpus-based typology.

Corpus-based typology aims at departing from the discretization of values for linguistic features and actually measuring them based on a corpus available for a language. There was a number of problems that had to be dealt with before making this methodology possible. Apart from lack of published

[1]By classical typology I have in mind works by Greenberg (1963), Dryer (1992) and alike.

corpora of documented languages (doculects), there first had to be invented a way to annotate the data consistently across languages, so that the corpora could be compared and the relevant information could be extracted automatically. The previously mentioned UD became popular precisely for that reason, as it involves a more or less theory-neutral way of annotating syntactic constituents and their relations, and is flexible enough to be extended to different types of languages. Not surprisingly, an increased number of linguistic studies (Levshina, 2019; Gerdes et al., 2021; Jing et al., 2021) started exploring this rich resource for different kinds of research questions, however typology is arguably the discipline that benefited the most from it, along with computational linguistics. Note though that the UD may be inappropriate or uneasy to use for certain research questions. For instance, it would be of little use for the studies in discourse typology. While one might use the UD with that goal in mind, the study would certainly be limited by the dataset and its annotation scheme in this case. For studies on discourse typology, Haig and Schnell (2021) developed a new annotation scheme called GRAID (grammatical relations and animacy in discourse) and applied it to a set of texts of typologically diverse languages. The increase in the number of published corpora collected during the fieldwork made possible the development of this more question specific dataset. Apart from that, another feature of the dataset is the use of natural speech, mostly in the form of traditional narratives, which distinguishes it from the UD. The latter mostly uses the written speech, which introduces a certain degree of bias in studies. Thus, in principle, if an appropriate annotation is developed,[2] many kinds of questions

[2]Note that this task is not a trivial one at all when we speak about cross-linguistic annotation.

can be researched in corpus-based framework.

1.2 Areal typology

As I have mentioned in the introduction, linguistic contact has long been recognized as a confounder in typological studies. As it is virtually impossible to collect detailed socio-historical information about all languages represented in a sample, researchers found a useful indirect measure for linguistic contact – geographical position. It comes as no surprise that languages that are close to one another are more likely to enter in contact and exchange some of their elements, be it grammatical features or lexicon, with the other languages spoken next to them. But due to lack of methodology, the researchers previously had no way to actually include this information in the study and they have thus resorted to the already mentioned sampling. All of that gave rise to a new concept of linguistic area (Hickey, 2017, p. 1–3):

> ... a region in which shared features among a number of languages are found with more than chance probability. The reason for such sharing lies in contact between speakers whose own language comes under the influence of others in their environment. Admittedly, *this view is simplistic*, but it is useful as a first approximation because it focuses attention on speaker contact. (My emphasis)

As Hickey highlights in the quote above, the notion of linguistic area is more of a convenience, rather than an accurate depiction of linguistic reality. The notion of geographical distance is continuous, whereas linguistic areas are typically discretized which looses a lot of resolution, especially when we get to the borders of a linguistic area. In more abstract terms, the problem is similar to that of discretization of linguistic features. The cause

for these decisions in both cases have already been mentioned and largely apply to the concept of linguistic areas too – the lack of relevant analysis methodology and difficulties related to data collection make it hard to resist the temptation of discretization. But it should be kept in mind that the development in statistical methods, Geographical Information Systems (GIS) and open linguistic databases make it possible to avoid these simplifications and achieve better inferences. I will discuss the methods to do that in the corresponding section. Despite these criticisms of the previous methodology, one should not expect a methodological revolution and instant integration of all the novelties, as researchers need time to get comfortable with the newest developments and adjust their study designs accordingly. In the following section, I will introduce a case that poses challenges for discretized use of linguistic features and linguistic areas.

2 Word order in Western Asia

2.1 Gradualness in linguistic areas

The topic of discrete linguistic areas is not a problem specific to the field of linguistic typology. In fact, linguistic dialectology has encountered the same problem long ago. Dialectology normally works with linguistic features specific to what is broadly referred as a language, which for our purposes can be defined as a set of mutually intelligible varieties, and then establishes what aspects of the language in question vary across geographically[3] distributed

[3]One might study the variation across different dimensions but normally variation across different social groups is studied by the field of sociolinguistics, whereas the variation across time is studied by historical linguistics. Needless to say, there is close cooperation

communities of speakers of this language. One of the main notions in linguistic dialectology is that of *isogloss* – a linguistic feature that is characteristic of a particular community's variety and is different from all/many other varieties of language in question (Ivić & Crystal, 2014). Dialectology has made significant advances in modelling the gradualness of change between different isoglosses (Jeszenszky et al., 2018), even though at the dawn of this discipline, researchers have been often drawing zones of different isoglosses on maps and essentially doing the same that typologists do when they talk about linguistic areas.

Jeszenszky et al. (2018) discuss the problem of isoglosses rarely having well defined boundaries but rather the switch from one isogloss to another one is gradual. Suppose we have two areas with two isoglosses. These areas will often have a center that would be a geographical place where an isogloss is most widely spread in the local community of speakers. The area between two such centers would then represent a transition zone where one feature is gradually declining in its frequency and the other one is increasing. The concept of transition is thus problematic for both classical dialectology and typology.

Finally, it is also needed to take into account historical contingencies of individual language families and of linguistic areas into account to avoid confounding in terms of linguistic features that languages inherit from their ancestor languages and also to avoid making generalizations confounded by language contact (Bickel et al., 2013).

In the absence of convincing evidence for functional-adaptive moti-

between the disciplines and works in these fields might even overlap.

> vations, I suggest that we accept that different types of syntactic constituents share their ordering patterns because they are historically related to each other, i.e. because they are linked by common ancestry. This also has important methodological consequences for typology. ... Just as other, more widely known, types of historical relatedness, such as a genealogical or areal interaction between two data points in a sample, need to be controlled for before one can test for a typological correlation, so does the language-internal historical relatedness between the grammatical patterns that make up that correlation. (Collins, 2018, p. 54)

Taking these historical contingencies into account can often help in explaining features that are typologically unusual and that are observed in languages. Provided we take the quote above seriously, one has to include grammatical features that might affect the relationship between features that we are seeking to explain but also phylogeny and contact relationship between languages in the sample. As I will be dealing with a transition zone in this study, all the aspects mentioned above are particularly relevant. Transition zones involve a geographical location where languages with diverse typological profiles come into contact and these languages undergo a unique mixture of changes both because of the fact that they belong to different linguistic families, but also because they enter in contact with typologically diverse languages. The intensity of contact along with languages' own internal innovations over time is likely to create languages with a lot of rarely found linguistic features.

2.2 Western Asia as a transition zone

Haig (2017) introduces Eastern Anatolia as an example of a transition zone in typology. To its West, in Europe, we have, with only few exceptions, VO and

prepositional languages, which are mostly represented by the Indo-European languages. To the East, one can find mostly OV and postpositional languages, including Indo-Aryan, East Iranian and Turkic languages. Finally, Semitic languages are spoken to the South of this area, most of which are either VSO or SVO and prepositional. An important aspect of the area is its great ethnolinguistic diversity. Armenian, Indo-Aryan, Kartvelian, Iranian, Semitic, Greek and Turkic speaking people, among others, still reside in the area. Haig (2017, p. 398–399) emphasizes another important socio-historical aspect, namely that the region was constantly dominated by different empires over the last two millennia and the speakers had to be multilingual to communicate with different ethnolinguistic groups. Current presence of few dominant languages (Turkish, Arabic, Persian) with typologically and genetically distinct profiles has further solidified Western Asia's status as a transition zone. Overall, the idea of transition zone appeared based on the previous research conducted in this area. For instance, the following quote is taken from a fundamental work on Iranian languages by Windfuhr (2009). Despite not using the term "transition zone", "buffer zone" is used to refer to a similar understanding of linguistic situation in the area:

> It is number six of eight isoglosses investigated in a succinct pioneering article by Stilo (2005), who also includes the relative position of demonstrative adjectives, numerals, adverb + adjective, object + verb, relative clause + noun, and object + adposition. Not only is the detailed distribution of this extensive set of isoglosses mapped within the Iranian speaking areas, but the Iranian isoglosses are also embedded in the wider context of the strictly right-branching typology of the languages to the west, represented by Semitic, and the strictly left-branching typology of the languages to the east of Iranian, represented by Turkic. Stilo could thus show how, overall, the multi-faceted internal dynamics of the Iranian languages reflects the mixed typologies

distinctive for a linguistic "buffer zone". (Windfuhr, 2009, p. 29)

The previously mentioned paper by Haig (2017) is a great overview of the area, while Haig and Khan (2018) is the key reference for a more detailed study of languages of the area. Some early ideas about transition zone, its properties and the difference between a transition zone and a linguistic area, can be found in Stilo (2005). There may exist other transition zones but references provided here and the examples from here on will apply to Western Asia.

In his study on word order of object and verb, Dryer (2013b) coded most of the languages of the area as being OV with the exception of Eastern Armenian. To see this plotted on a map, please see Figure 1. The map additionally reflects the position of Western Asia as a buffer/transition zone. Given that Dryer's study was more focused on the distribution in the languages of the world, it comes as no surprise that many minor languages were not included in the study. All of that said, the main point I want to make is that Dryer's map creates a biased impression of the languages of the area as being OV and hence being simply an extension of general Central and South Asian tendency towards being OV. Nonetheless, if the combination of this feature and of feature 85A, 'Order of Adposition and Noun Phrase' (Dryer, 2013a), is taken into account, one can observe strange pattern of some languages in the area being OV and prepositional (e.g. Central Kurdish, Jewish Neo-Aramaic, Persian) which is at odds with Greenberg's Universal 4, which states that languages with OV order tend to be postpositional.[4] This raises a reason-

[4]The universal mentioned here is, in fact, non-categorical. Greenberg (1963) uses the following wording: "with overwhelmingly greater than chance frequency". Nonetheless, this tendency is strong and the cases mentioned in the paper is worth studying.

able question about the details of word order in languages of Western Asia. Given the linguistic diversity and the previously mentioned transitional position of the area, we might expect to find some heterogeneity in word order, if corpus-based method is applied.

2.3 VO and OV asymmetry in the oblique order

An additional interest in this area and non-subject constituents' order is explained by an empirical observation made by Dryer and Gensler (2013). I will first introduce the observations made in this study and in Hawkins (2008), and then explain in more details the relevance of it for Western Asian languages.

The dataset that Hawkins used contains 500 data points about the dominant word order of object, verb and oblique. In this dataset, oblique is defined as "a noun phrase or adpositional phrase (prepositional or postpositional) that functions as an adverbial modifier (or "adjunct") of the verb" (Dryer & Gensler, 2013). The data is reproduced here for convenience in Table 1. The important observation is that in VO languages (first two lines from the table), the position of the oblique is almost always fixed at the last position. In fact, the 3 languages that have XVO order are all Sinitic and spoken in modern China. VXO word order is not attested at all. At the same time, OV languages have diverse positions for obliques.

Hawkins (2008) studies this observation in greater details to find an explanation for this asymmetry. His explanation is based on processing efficiency. There are several patterns that were proposed by Hawkins (2008, p. 188) and also two general principles that govern the linear order of constituents.

Figure 1: Verb and object positions in languages of Asia, Europe and North Africa (Dryer 2013b).

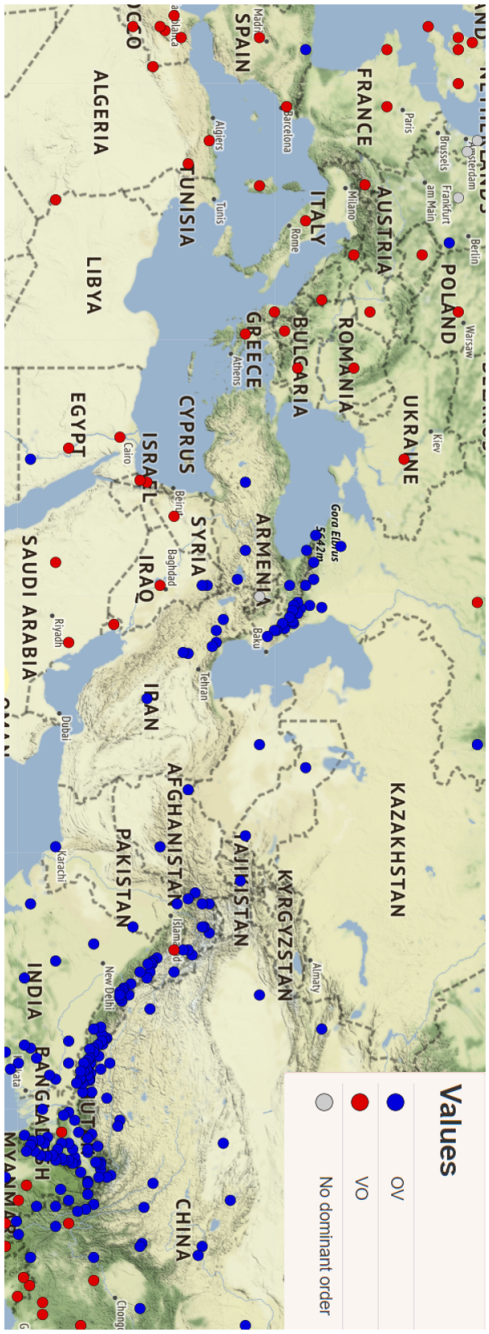


Table 1: Order of Object (O), Oblique (X), and Verb (V) in Dryer and Gensler (2013).

Value	**Count**
VOX	210
XVO	3
XOV	48
OXV	27
OVX	45
No dominant order	167

Overall, the predicted relative frequency of word order types is: VOX > XOV > XVO/VXO. Another explanation, similar on the surface but different in its details, is described and tested in Futrell et al. (2015). The study by Futrell et al. (2015), in essence, tests the hypothesis whether languages have dependency length[5] less than it would have been if dependency lengths were distributed at random. Note that the latter study has the advantage of using Universal Dependencies and thus testing the hypothesis on multilingual corpora, rather than using discretized features, as in the study by Hawkins (2008).

The implications of processing typology will not be tested here and were presented to introduce one view on the explanation of word orders. In the present study I would like to touch on a different explanation of word order, namely the one that accounts for genetic inheritance of languages and of language contact, and their interaction. Importantly, the study will try to provide one possible methodological solution to deal with the problems

[5]Dependency length is a measure that indicates the distance, in syntactic words, between a node and its dependent elements. For example, in UD dataset, the distance of object and its head in 'I have a car' is two because there is a determiner between the object 'car' and its head 'have'.

outlined in the first two sections. Testing the impact of phylogenetic and contact history is particularly important in case of transition zones, as we would expect more intralinguistic heterogeneity in such linguistic areas. As I will try to show in this study, the latter seems to be the case.[6] Overall, this goes in line with the research philosophy of language typology that is proposed in Bickel et al. (2013). Apart from that, I will also try to make some generalizations and test some claims about this area and its representative typological features that were proposed in Haig (2017).

3 Methods and data

3.1 WOWA dataset

The Word Order in Western Asia (WOWA) corpus (Haig et al., 2021) is a project developed at the university of Bamberg that is designated to serve for various theoretical and methodological goals. Out of 25 doculects available at the moment of running the analysis, I excluded one doculect and ultimately had 24 doculects. Christian variety of Neo-Aramaic spoken in Barwar was excluded from the analysis, as it did not have geographical coordinates and was difficult to be mapped to any of the existing varieties of Christian Neo-Aramaic found in Glottolog (Hammarström et al., 2021). The latter will turn out to be important further in the paper.

The corpus consists of transcribed spoken data, most of it collected by linguists during the fieldwork. The utterances are separated into what roughly

[6]It is hard to test this claim more generally because there is no comparable dataset of this size and designed with the same purpose as WOWA dataset, which will be introduced in the following section.

Table 2: An example of two WOWA data points.

affiliation1	affiliation2	doculect	token	token_translation	pro	anim	weight	role	flag	position	nc	comments
hell	ponticgreek	romeyka	emas	us	1	hum	1	do-def	case	0		
iran	gorani	gawraju	dasim	my hand		bp	1	goal-c	bare	1		directional particle on the verb

corresponds to a clause and each clause constitutes a data point. Only clauses that contain non-subject constituents are coded. Moreover, an element has to be referential to be considered. The clauses that contain one or more non-subject elements, e.g. obliques and objects, are then coded in accordance with the coding guidelines found on the official website of the corpus. If a clause does not contain non-subject elements or they are inappropriate for coding,[7] the data point has value '1' in the column 'nc'. Otherwise, the column 'position' represents the variable of interest and is coded in a binary way, having values of '0' or '1', depending on the position of the non-subject element. The values correspond to being pre-verbal or post-verbal element, respectively.

Furthermore, each codable data point is annotated for a series of variables, each corresponding to different linguistic and phylo-geographical features. In Table 2, an example of two coded data points from WOWA dataset is shown. I removed some of the columns, such as geographical coordinates and the clauses that contains the tokens to fit the table.

The variables that are worth commenting include columns 'weight', 'role' and 'flag', as these will represent the main structural predictor variables in the model introduced in the following subsection. The variables' coding will be simplified compared to the original coding scheme to avoid too many pa-

[7]For example, Semitic attributive copula clauses in present tense were considered non-codable, as they lack an overtly expressed copula verb.

rameters for some of the variables. 'Role' is one the central linguistic features used as a predictor. Our main interest is about the position of direct objects (coded as 'do' and 'do-def' in the dataset) in comparison with what Hawkins (2008) called oblique elements. To be able to test some of the hypotheses presented in Haig (2017) I further introduced the following categories: goals, non-subject arguments of copula verbs, and non-subject arguments of the verb 'become'/'turn into'. All the other elements that do not correspond to any of the defined categories were merged into the category 'other'. The variable 'flag' is about the grammatical or adpositional marking of a non-subject element. There are various ways in which languages may mark these elements. The first one is just by using word order and it lacks any additional elements. In this case, flagging is said to be 'bare'. As opposed to that, some languages have case marking or could use adpositions to mark certain elements. For instance, English oblique object in ditransitive clause 'John gives a book to his elder brother' is marked through preposition. All of the cases where any kind of flagging, apart from word order, is found, will be called 'non-bare' or 'flagged'. The interest for this variable is a long-standing one. Kiparsky (1996) proposed a famous hypothesis that the Germanic languages that lost inflectional case marking are less variable in their word order. This hypothesis is currently under debate but given the specifics of the model I will introduce, it will be possible to test whether the hypothesis may be extended to the languages of Western Asia, i.e. we expect more variability in word order of non-bare elements compared to bare ones. Finally, 'weight' represents the number of non-bounded (non-affixes and non-clitic) words that an element consists of. Note that the value of 4 means that an element has 4 or

more words. This variable should supposedly increase the probability of the element being shifted to post-verbal position, in accordance with Heavy-NP shift hypothesis (Arnold et al., 2000).

3.2 Methods

As discussed in the introduction, there is a variety of methods that were used to control for phylogenetic inheritance and language contact. The model I want to focus here is presented in Jing et al. (2021) and outline some of the conceptual problems with its approach to the control for phylogenetic influence. The model that Jing et al. (2021) use is a hierarchical[8] one. This kind of models assumes that the categories of a categorical variable come from a single population and although the categories have numerically different effects on the outcome, they are constrained by a hyperparameter that defines a distribution for the whole population (Gelman et al., 2013, p. 101–132). Hyperparameter often corresponds to the mean and standard deviation of a normal distribution. For a conceptual example related to linguistics, suppose we have a single language family and the observations in the dataset are coded as being clades of that language family in the form of categories. For instance, if we were to pick languages from the Indo-European language family (IE), we could have categories such as 'Germanic', 'Slavic' and so on. In this case, hyperparameter describes the distribution of the effects for the entire IE family, whereas each individual effect for 'Germanic', 'Slavic' and all the other pre-defined clades may differ but they still come from the same distribution that is described by the common IE hyperparameter. In

[8]Also often called *random effects* model or *multilevel* model.

this way, all clades inform the shape of the distribution but the hyperparameter constraints individual effects of clades. These models have numerous advantages but for our purposes, they also have two problems:

- The model assumes that all the categories come from a single distribution and that all the categories share some information because of that. Previously cited work by Jing et al. (2021) uses languages that belong to different unrelated language families and it appears to be inappropriate to apply the hierarchical model in this case, despite the better fit that it provides.[9]

- The observations that belong to a single category, e.g. 'Germanic', will all have the same effect on the outcome. But Germanic language family, despite being only a clade of the IE, is still a big subfamily with a lot of daughter languages. And some of those daughter languages are more closely related than others. Danish and Swedish are both North Germanic languages and we would expect them to have more similarities than any of the two has with High German. Thus, a typical hierarchical model will depend a lot on the specific level that we use for the categorization.

Naranjo and Becker (2021) use instead a Gaussian process model to control for both phylogenetic inheritance and for possible contact influences between languages. This is also the approach that I am going to use in this paper.

[9]One might argue though that the strength of the effect will be similar across different families. That is, if we learn how much influence belonging to a family affects a certain feature, we might assume that the strength of this effect would have been roughly the same in a different language family, given everything else being equal.

Gaussian process (GP) models provide a way to represent the continuity in covariance between closely related languages, and less covariance in the more distantly related languages. Numerically this means that languages that have low distance between them, their effects are going to be similar. The notion of distance is of great importance for GP models and I will explain in more details the way I computed the relevant distances for the doculects represented in WOWA. To compute the covariance between any pair of languages, a kernel is used (Gelman et al., 2013, p. 503). Kernel is a function that defines how the increase/decrease in distance affects the covariance between a pair of doculects. There exist a variety of kernels used in different fields for different kinds of distances.[10] After taking into account how effects for related doculects covary, Gaussian process ultimately results in an intercept which will be used to make predictions. As the model will account for some structural features as well, phylogenetic intercept should roughly show the general tendency of a language being pre-verbal or post-verbal, if we leave out the rest of the factors. Contact intercept will instead show the effect of being pre-verbal or post-verbal that is not explained by phylogeny and judged likely to be the effect of language contact. The posterior distribution should ideally reflect comparative historical linguists' judgements about the languages' genetic and contact history. For example, Proto-Semitic and the absolute majority of modern Semitic languages have their object placed after the verb (Huehnergard, 2019, p. 68–69). Thus, the phylogenetic intercept should have a value strongly predicting consistent post-verbal position of constituents, while contact influence could be pooling the word order to-

[10]Impressionistically, people tend to think of physical distances only but one may define and compute, e.g. cultural distance. Such distance would apparently represent a construct.

wards being sometimes pre-verbal, as e.g. it is the case in some Semitic languages spoken in Western Asia.

While I will describe the kernels used for the model, it is also necessary to mention the distances that will be used in the Gaussian processes. After all, the distances I used are not an inherent component of the WOWA corpus but rather are computed based on the data available in WOWA.

I will start with a conceptually simpler distance, that of geographical proximity. It is evident that geographical distance does not have any direct influence on linguistic contact and that contact overall has many layers that need to be accounted for, when thinking about a language borrowing lexicon or grammatical features from another language. But the whole information across all the layers is typically unavailable. The information about trade relationships, social structure, prestige of a variety, and many other important variables in language contact are sometimes unknown even for synchronic varieties. Needless to say there is less information with regard to the whole history of language contact. At the same time, geographical distance is a useful approximation of the degree of language contact. Geographically distant languages, even with all the other conditions for borrowing being fulfilled, are not expected to have a large number of loaned structures and lexicon between them. That being said, the geographical distance between two doculects is not a simple matter. First of all, languages are spread in space and it is hard to delimit the boundaries. In case of this study, the advantage we have is that the descriptions found in WOWA have a well documented place of recording and place of socialization of the speakers who contributed to the corpora available in WOWA. The second issue is related to the kind of dis-

tance we want to use. For example, Naranjo and Becker (2021) used the shortest path between any two places where a language is spoken. This distance works well for datasets where languages are distributed around the globe and there are few languages per area. In our case, we have a delimited area and would like to use a more realistic distance to have finer details about what it would take speakers of a doculect to travel to get in contact with speakers of another doculect. To do that, I decided to use Google Maps[11] walking distance. Walking distance reflects the fact that there are certain geographical objects, e.g. mountains, that are to hard to pass through directly and speakers would have to pass them through some indirect paths. Conveniently, Google Maps takes into account the fact that for certain routes, it is easier to take a boat trip and thus reduces the travelling distance drastically. For instance, Kumzari speakers could cross the Persian Gulf to get in contact with Balochi speakers instead of passing through the entire Arabian peninsula. For obvious reasons, the algorithm considers modern roads and pathways that could be used for walking. One might argue that these roads are modern and might have not existed in the ancient times. While the objection represents a valid concern, many modern roads are built and represent ancient routes of people's movement, as roads are typically built when they are demanded. Additionally, there are two technical moments I have to mention. Due to Google Maps' algorithm peculiarity, the distance is not always entirely symmetric. Because of computational requirements, I manually made all the distances symmetric. The model will infer covariance matrices and they have to be symmetrical. Another change introduced

[11] I used R package *gmapsdistance*.

by me, is slightly shifting the Northern Kurdish variety of Ankara to avoid complete overlap with Turkish variety of Ankara. Zero distance off the main diagonal of distance matrix might again produce computational instability but I made sure the distance remains low enough to represent a high degree of contact in Ankara. The final distance matrix used in the model, along with all the other materials and code can be found in a GitHub repository.[12]

The kernel I will use for geographical distance is widely applied to distance measures (McElreath, 2020, p. 469). It uses squared distance between two points in its formula. As the distance is squared, it means that contact intensity will decay more rapidly as the distance becomes large. I will later compare this kernel with the one used for phylogenetic distance to explain why it is conceptually convenient to have it this way and also outline reasons for using two different kernels. The full formula of the kernel for geographical distance can be seen below. The description of what each parameters correspond to is taken from McElreath (2020, p. 469). The squared distance here is D_{ij}^2. Please note that the kernel for phylogenetic distance is going to be similar in its formula, the major difference consisting in non-squared distance.

$$\mathrm{K}_{geo[i,j]} = \eta^2 \exp(-\rho^2 \mathrm{D}_{ij}^2) + \delta_{ij}\sigma^2 \tag{1}$$

Since the two kernels are identical for the most part of it, I will describe what each parameter above means. The first parameter, η^2, represents the maximum covariance between any pair of doculects. It is multiplied by the term $\exp(-\rho^2 \mathrm{D}_{ij}^2)$. Note that the parameter ρ is negative, which means that

[12] `https://github.com/acraevschi/Bayesian-WOWA-Model`

the higher ρ is, the faster is the decline. An interesting use of it might be found when GP is used for two different features in two different models. Afterwards, one could compare what the model estimates for ρ are to check if any of the features is declining more rapidly than the other, based on model's estimations. Finally, the last two terms are estimated as a single parameter in the model and account for some variability in covariance beyond the first two terms. This parameter can be fixed at constant value but this would take some flexibility from the model, so I decided to constrain it to be very small in case of both geographical and phylogenetic distance. See Appendix 5 or the Stan code in GitHub repository for prior distributions but what is conceptually happening in the model, I set phylogenetic distance to have little variability, i.e. $\delta_{i,j}\sigma^2$ is low, while the same term for geographical distance is tolerated more to have a higher value. If isolated, we expect two languages to change at a roughly equal rate,[13] while geographical distance might or might not impede the contact. Often times there are other factors beyond geography that affect the intensity of linguistic contact and since the model only has geographical distance, we could allow it to have some more variability when inferring this last parameter.

Phylogenetic distance between two languages is a more subtle case and there will be some arbitrary decisions made about it in this work. The reason for this arbitrariness is lack of research and rare use of phylogenetic regression in linguistics. Phylogenetic distance represents the time of divergence of two taxa (doculects in this case) from their most recent common ances-

[13]It is important to distinguish the quality of change from the quantity of change. I.e., two languages might go in completely different directions of change but if they do so, we would expect the rates of change to be about equal. This assumption needs to be verified but it is assumed in the model.

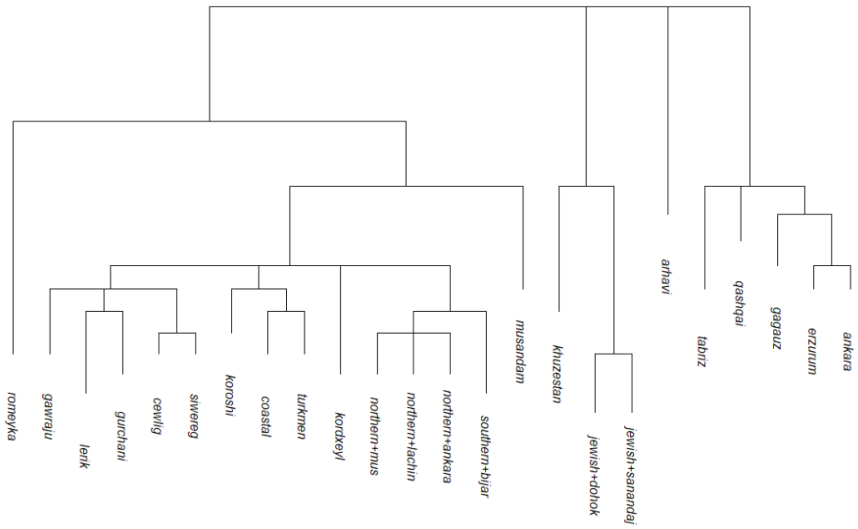
tor. The time is expressed in different ways, depending on the phylogeny. Measure of time is proportional to branch lengths of a phylogenetic tree. If phylogeny is calibrated, this time corresponds to the number of years back in time when the divergence occurred. However, in case of this work I will use a non-calibrated phylogeny – a cladogram. Cladogram is a tree that contains topology[14] but branch lengths are missing or set to an irrelevant constant. As explained later, branch lengths will be transformed. In case of WOWA corpus, we have at least 4 major families and for convenience I decided to use a more or less neutral phylogeny that one can find in Glottolog (Hammarström et al., 2021). Glottolog's phylogenies are neutral in that they most often correspond to the majority opinion of the scientific community. This may make it conservative, for example, disputed Transeurasian language family (Robbeets et al., 2021) is not represented in Glottolog, but since we are dealing with relatively well studied language families, we could accept Glottolog's phylogeny as at least a good approximation of real tree's topology.

To extract phylogenies, first I identified the doculects from WOWA and assigned to them Glottocodes. Some of the WOWA varieties are not yet represented in Glottolog. To account for that, I have assigned these doculects the variety that should be a close sister variety to it based on the description of a WOWA doculect. This procedure of assigning not exact varieties glottocodes to a doculect is harmless, as we are only interested in the relative position of a doculect. The glottocodes that were assigned to different doculects can be found in the previously mentioned GitHub repository. To

[14]The relative positions of taxa in a tree.

extract the phylogeny, I used an R package *glottoTrees* by Round (2021). The ultimate tree is displayed in Figure 2.

Figure 2: WOWA tree used in the model.



The trees extracted from Glottolog initially have an equal length but given what we see in other phylolinguistic studies, this is not a realistic assumption. For that purpose, the branch lengths were rescaled with the help of *geiger* package by Harmon (2020). As there are no studies on which scaling is most appropriate in case of language phylogenies, I compared various types of scalings used in phylogenetics. One was exponential, proposed by Macklin-Cordes and Round (2022). This is based on phylogenetic studies conducted in evolutionary biology. This scaling exponentiates the length of each branch above the tips and further exponentiates the branch lengths of the subsequent branch lengths, hence makes the recent evolution of languages relatively slow. To understand the extent of it, if branch lengths are normalized (transformed to interval from 0 to 1), this scaling assumes complete identity of closely

related varieties, e.g. of Northern Kurdish varieties.[15] To try a less radical scaling, I tried other options, among them δ scaling, scaling with κ and scaling with λ, all of them taken from Pagel (1999). Harmon (2020) describes δ scaling in the following way: "Where δ is greater than 1, recent evolution has been relatively fast; if δ is less than 1, recent evolution has been comparatively slow. Interpreted as a tree transformation, the model raises all node depths to an estimated power (δ)".[16] δ-values that I tried are (1) δ=1.5, and (2) δ=0.65.

Another kind of scaling that was used is κ transformation. This transformation scales branch lengths in dependence with the number of speciation events. The two values were picked arbitrary, they are: (1) κ=1.5, (2) κ=0.5.

The last kind of scaling that was applied is λ transformation. λ can takes values between 0 and 1. Values closer to 0, make tree more star-like, which means that more uncertainty is introduced and distance between more closely related varieties gets bigger, while distances between unrelated languages remain unchanged. λ=1 would make no difference to the tree, so I decided to try λ=0.5.

It would be useful to empirically verify, which scaling would work as a better heuristic in linguistic studies that involve phylogenetics, this would be an interesting topic for an eventual study. I ran five identical models with the only difference being the scaling of the trees to compare which of the scalings is a better heuristic, at least in case of word order. The normalized phylogenetic distances between doculects for scaling with δ=0.65 is shown in

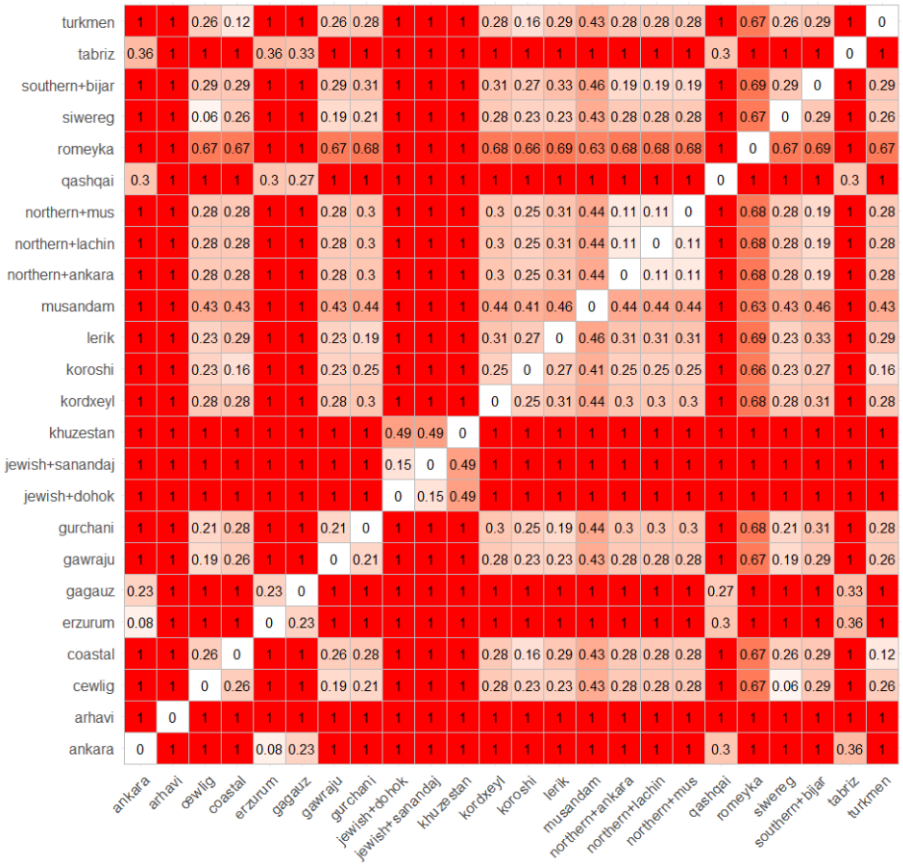
[15] In fact, correlation between these closely related varieties only converges to 1, which is the value of complete identity, but due to rounding error, computer outputs 1.

[16] The quote is taken from the documentation of *geiger* package.

Figure 3. This scaling will turn out to work best, as I will describe in the next section.

Figure 3: Normalized phylogenetic distance between WOWA doculects.

(a) Distance of 1 corresponds to doculects being unrelated.



The kernel used for phylogenetic distance differs from the one used for geographical distance. This kernel represents what is called Ornstein-Uhlenbeck process (OU). This kernel differs in that doculects would still covary between them even when phylogenetic distance grows large. Thus, word order of related doculects is assumed to preserve some traces of common ancestor even

after a relatively long period of time. This assumption is confirmed by Dunn et al. (2011). The formula for the kernel is identical to the Equation 1 but this time the distance is not squared:

$$\mathrm{K}_{phylo[i,j]} = \eta^2 \exp(-\rho^2 \mathrm{D}_{i,j}) + \delta_{i,j}\sigma^2 \tag{2}$$

In order to account for some variability that might be found between the speakers, I used 'textID' column. This is not entirely correct, as some of the texts are pronounced by the same speaker but this information is not available in all the cases. As the model is Bayesian, I have used prior distribution to limit the influence of individual variability, as we do not expect a doculect to differ significantly between individuals from the same, relatively small area that corresponds to the place of socialization of speakers.

Additionally, I have introduced an interaction term between 'role' and 'flag' variables. If we assume that presence of flagging increases the liberty that a speaker has in choosing the position for an element, there might still be differences in the extent to which it applies to different grammatical roles. Some of the roles may be strictly fixed in their word order, be they flagged or not, whereas some may instead acquire flagging only to get shifted to a non-standard position. This might happen for pragmatic reasons, for instance. In statistics, when interaction term is mentioned, it normally takes form of one variable (Z) being multiplied by the other (X) plus including both variables separately in the model, as in the following example:

$$Y = X \cdot Z + X + Z \tag{3}$$

This form assumes that both variables, X and Z in the example above, have some effect on the outcome Y separately but their interaction $X \cdot Z$ is also important. In case of 'role' and 'flag', we could ask a question of whether flagging has any effect of its own on the position of an element. But this does not seem to be the case. Assuming we know that an element is flagged with preposition, it provides very limited information about the position of this element and overall it does not appear to have a major impact of its own. Because of that, I decided to use a different kind of interaction term, exemplified in Equation 4:

$$Y = X \cdot Z \tag{4}$$

This means that 'role' and 'flag' are important only in their interaction, that is, every combination of 'role' and 'flag' might have a different effect on the outcome. This would allow us to say which grammatical roles have more flexibility in their word order when they are flagged in some way and maybe some of them will not differ and will be more restricted in their distribution in a clause.

I decided not to use animacy as predictor variable as it is to some extent caused by grammatical role or the other way around. Knowing grammatical role increases the probability of element having a certain animacy value. This might create a mediator effect, for more on it, check Cinelli et al. (2020).

Finally, in Equation 5 the full model is presented. Variables P and G correspond to the above described Gaussian processes, phylogenetic and geographical distances respectively. All the variables, apart from 'textID' (T) were assigned flat priors. That means that I introduce no prior beliefs about

the strength of the effects of any of the variables. It is acceptable and works in this case, as the dataset is large. In case of T_i, prior information constraints the model on sampling values around zero for the variable. The outcome variable 'position' takes the values of 0 and 1, which correspond to 'pre-verbal' and 'post-verbal', respectively.

To predict this binary outcome from a set of predictor variables, binomial logistic regression models are used. Logistic regression model predicts the outcome in terms of a parameter p, which corresponds to the probability of having post-verbal element position (i.e. position=1). I used logit as a link function, which is a standard choice in many occasions. For more details and peculiarities of logistic regression models, see Gelman et al. (2013, p. 405–431) and McElreath (2020, p. 323–365). I randomly divided the data in two chunks, 70% and 30% corresponding to training and testing parts, respectively. Thus, I will only use the first chunk for the initial inference of parameter values and the second to test how well the estimated parameter values predict the unseen data. Another important point I would like to emphasize is that the model almost lacks language specific parameters. Doculect's geographical position and its position in the tree are the only two parameters that are language specific but even they define a language in somewhat structuralist sense, i.e. in terms of its phylogeographic position compared to other doculects. Thus, we could potentially introduce a new language in the dataset, with the condition that it is coded in accordance with WOWA guidelines, locate it on the tree and geographically, and then predict its non-subject elements' position with the model.

$$
\begin{aligned}
position_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= P_i + G_i + RF_{[role_i, flag_i]} + T_i + \beta_w W_i
\end{aligned}
\tag{5}
$$

Apart from variables described above, $RF_{[role_i, flag_i]}$ stands for an interaction of 'role' and 'flag' of the data point i and W is 'weight'.

To fit the model, I used Hamiltonian version of the Markov Chain Monte Carlo (MCMC) algorithm as implemented in Stan software (2022). The model ran for 3000 iterations, half of which were discarded as warmup and the rest was used for sampling of the posterior distribution. MCMC is a stochastic algorithm and requires some additional care before using the results it produces. For instance, a basic recommendation is to use multiple chains sampling in parallel to ensure that a MCMC chain actually found a correct posterior distribution, rather than an accidental one. There were 4 chains running in parallel to ensure that the chains converge and that the posterior distribution is explored sufficiently well. To improve sampling, non-centred parametrization is used but it is not reflected in Equation 6 from Appendix 1, as it would take additional space, while it has no effect on the posterior distributions of parameters. The non-centered parametrization was done in accordance with Stan's recommendations (Stan Development Team, 2022) and the exact implementation can be seen in Stan's code in GitHub. The goal of non-centered parametrization is one of improving the

performance and efficiency of MCMC sampling.

4 Results and discussion

4.1 Model comparison

Before introducing the results, such as posterior distributions of the parameters and plotting model's predictions, I would like to compare various models in terms of their performance and its implications for the interpretation of model's results. Running multiple models has long been considered a kind of scam from the researcher's side, as this allows to formulate theory based on the single best model, while scientific workflow requires to proceed in the direction where theory informs the model. Nonetheless, as McElreath (2020, p. 5–7) points out, sometimes models are compatible with various theories and because of that, it does not make sense to select a single model until more aspects of empirical observations are in one way or the other accounted for by the model. Final argument against comparing multiple models is that this approach is often associated with a particular type of scientific misconduct called p-hacking. Note that the latter is associated with a particular framework of testing hypotheses, when researcher selects a null hypothesis, e.g. that grammatical role has not effect on the position of a token in WOWA. After running the model, researcher would then test whether any of the roles has significant effect on the position of a token. The obvious problem with this approach is that it is too simplistic, especially in case of language science which studies a complex system like language. It forces one to compare a model in which grammatical role has no effect on the element's position to a

model where it does, while every linguist would agree that the former model is absolutely unrealistic. Because of that, I proposed two models of interaction between grammatical role and flagging of an element. In one case, flagging has an effect of its own combined with interaction between flagging and grammatical role, whereas in the other model flagging only matters in terms of its interaction with role. For the sake of completeness, these two models will also be compared to a model which has no interaction.

In this paper I will use procedure called Widely Applicable Information Criterion (WAIC) to compare the models. WAIC approximates model's out-of-sample predictive potential. This is done by procedure called cross-validation. This procedure removes a certain part of the data and estimates model's performance on the removed data. Then this performance is quantified. This yields a kind of "performance" score of the model that we can compare. An important nuance is that WAIC also penalizes additional complexity of the model, that is, whenever an additional parameter is added, this represents additional complexity. This makes sense, as mathematically it is easier to fit the model when the number of parameters is high and because of that, if the number of parameters increases, predictive performance should increase significantly to justify the added complexity and avoid overfitting.

Based on the results from the Table 3, there is no big difference between the models. WAIC score for the first model is marginally better, as it is lower but overall the difference is negligible, especially if we take into consideration standard error of WAIC. What might have affected WAIC in preferring the first model is a lower penalty term. Later in the work I will nonetheless use the first model as it appears to match the theoretical knowledge better and

Table 3: Model comparison via WAIC of three implementations of the model.

(a) **WAIC** - WAIC score

(b) **SE** - standard error of the estimated score

(c) **ΔWAIC** - difference in WAIC scores from the best

(d) **pWAIC** - penalty term.

Model	**WAIC**	**SE**	**ΔWAIC**	**pWAIC**
role · flag	8131.8	131.11	0.0	94.9
role · flag + + role + flag	8133.2	131.16	1.4	95.5
role + flag	8133.3	131.09	1.5	95.4

has a marginally better WAIC score. It does not seem likely that flagging on its own has any impact on the position of an element.

There is no big difference between different kinds of scaling and for the sake of space, their comparison is not shown here. The scaling with $\kappa = 1.5$ and with $\lambda = 0.5$ was somewhat worse than the rest, while the marginal leader is $\delta = 0.65$. Most likely, since the parameters have a lot of freedom to vary, the difference in scalings is compensated by this freedom and ultimately MCMC could find optimal values for η and ρ to produce more or less the same results. In the end, the scaling with $\delta = 0.65$ was picked as the model whose results will be presented in the following section. Another nuance that needs to be mentioned is correlation between unrelated languages. For theoretical reasons, we do not assume any relationship between languages that are classified as belonging to different macrofamilies, while the model finds small correlation between them. To avoid this theoretical inconsistency in modelling, I decided to further manually scale the distance between unrelated languages by a factor of three. That means that unrelated languages

had a distance of 1 between them and after this transformation was applied, the distance equals 3. This factor has no justification in literature but what it achieves is that correlation between unrelated languages is now equal to 0, which is more theoretically sound. Note that the latter model has a slightly worse WAIC score[17] but this small fitting sacrifice allows to better account for theory.

The best option to select an appropriate scaling would ultimately be inferring it as a parameter while sampling, i.e. the solution is to include unscaled Glottolog distances to the model and make the scaling factor to be a parameter. This implementation would potentially provide a better fit but is costly in terms of computational power.

4.2 Results

After inspecting Stan's main statistics about the chains' convergence, $\hat{R}4$ and the number of independent samples, the model appears to sample well. It is desirable to have $\hat{R}4$ converging to the value of 1.00 for all the estimated parameters, although there is no accepted norm. For instance, Naranjo and Becker (2021) consider values as high as 1.04 acceptable. There is no standard minimum for the number of independent samples as well. McElreath (2020, p. 281) suggests that even 200 samples might be enough to explore the posterior distribution. In case of this model, all the parameters had more than 1220 independent samples with the mean value of 5860 independent samples per parameter. Model's $\hat{R}4$ values were also close to 1.00, all of

[17]The model that has unrelated languages' phylogenetic distance scaled by a factor of three has WAIC score that is worse by 0.7. This difference is low enough to not impact the predictive quality of the model significantly.

them being lower than 1.01. Visual inspection of the chains via traceplots also suggests good mixing of chains. Due to stochastic nature of MCMC algorithm, one needs to assure that all the chains sample roughly the same values. All of this points to the model reliably sampling the posterior distribution of the parameters. At the same time, some parameters have more uncertainty, which can be seen in the shape of the posterior distribution. As will be shown later in this section, the most uncertain parameters are the ones related to Gaussian processes, that means that the model cannot always find the most optimal contact and phylogenetic effect strength.

I will display and describe the full posterior distribution of the parameter values in the text that follows. In Bayesian statistics, the model infers not a single estimate of the best parameter value but rather it produces an entire distribution of the probable parameter values. Thus, one should focus on the shape of the distribution of a parameter and on the values that are included in this distribution. The distribution shows us the probability that a parameter has a particular value given the data, mathematically expressed as $\mathrm{P}(H|D)$. This is a different interpretation from that found in frequentist statistics and should be interpreted in accordance with Bayesian definition of the interval. Because of this peculiarity of Bayesian models, I will either introduce the confidence interval for a given parameter, or plot the distribution in an appropriate form. 95% confidence interval (CI) is a standard confidence interval range for both frequentist and Bayesian statistics. I will use this range hereafter. It means that with 95% chance, the true value of a parameter is in the reported range.[18] Because of the mathematical properties

[18]Because of the properties of MCMC, one can say that the algorithm considers all the models, i.e. all the possible parameter values and their combinations, and samples them

of logistic regression, most of the values were transformed by applying the inverse logit function. Inverse logit transforms the value of a parameter to map it to an interval from 0 to 1 to reflect the effect of a variable on the outcome on the probability scale. Hence, if a parameter has values closer to 1, it is more likely that an element is going to be post-verbal and the reverse is true. At the same time, it also makes sense to look at log-odds values of some of the parameters, as this allows to decide on whether a particular variable has any significant effect on the outcome. For the sake of space, I will often show posterior distributions of parameter values on either probability scale or on the log odds scale. If posterior distribution only ranges from 0 to 1 on one of the axis, that means that probability scale is used. When distribution includes negative and positive values, that means that log odds scale is used.

I will start with posterior distributions of different combinations of roles and flags plotted on the Figure 4. We are most interested in (1) the distribution of roles' positions, and (2) comparison of posterior distributions' shapes of flagged versus non-flagged roles. If posterior distribution is wide, the model is less confident about the parameter value and this serves as a proof to the hypothesis about the influence of flagging on constituent position. For instance, visual inspection of Figure 4 allows us to say that the position of flagged direct objects 'do-non' is less certain than the position of non-flagged direct objects 'do-bare'. The same appears to hold, at least to some extent, in case of goals and copula arguments. There is an additional figure of these three roles in Figure 5 that compares their distributions. Conversely, the distribution of arguments of 'become/turn into' and

in proportion to their representation in the posterior distribution.

of the category 'other' are generally wide and there is no major difference between bare and flagged elements. Overall, it appears that some roles are indeed more variable in their position, as Kiparsky (1996) suggested. This interpretation arises from wider posterior distribution, as shown in Figure 5. Note though that there are multiple factors that could explain uncertainty in the estimated posterior distributions of flagged elements. Importantly, the distribution of flagged and bare constituents in WOWA is almost identical: 46% of coded constituents are bare and 54% are flagged. Thus, it is unlikely that the number of observations per category influenced the shape of the distribution that was inferred. Despite this fact, I will avoid categorical yes/no answer to hypothesis by Kiparsky (1996) and will limit myself by saying that posterior distribution seems to support it but only for some roles.

In Figure 6, mean values for the effects of the variable 'TextID' are shown. In this Figure, values on log-odds scale are displayed to appreciate to what extent the individual variation contributes to selecting the position of an element. From that Figure, one can see that in most of the cases, individual variation has little influence on the decision between pre-verbal and post-verbal position of a non-subject element. Not surprisingly, the model attributes the strongest effects of individual variation to the doculects in which it is difficult to observe a consistent pattern in word order. For example, one of the strongest mean effects of individual variation (-1.20 on log odds scale) is found in one of Romeyka's texts. Schreiber (2018, p. 921) mentions that there is still no consensus as to whether Romeyka is predominantly VO or OV, as there is a lot of internal variation attributed to different sociolinguistic factors. Inspection of Romeyka corpus' metadata shows that

Figure 4: 95% confidence intervals of different roles and flags combinations.

(a) Points indicate the mean value of the posterior distribution of a parameter.

(b) Higher values of p mean that the combination of role and flag are likely to be post-verbal.

(c) Values in the middle represent roles and flags' combinations that have no particular preference in their position.

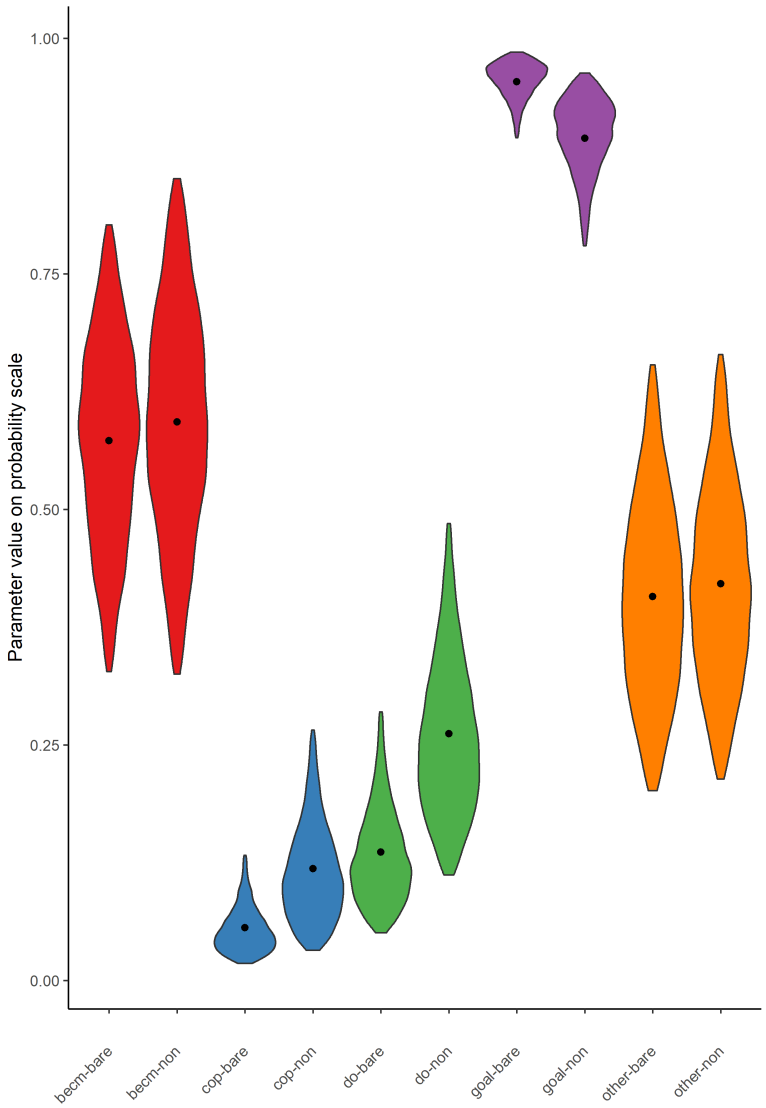
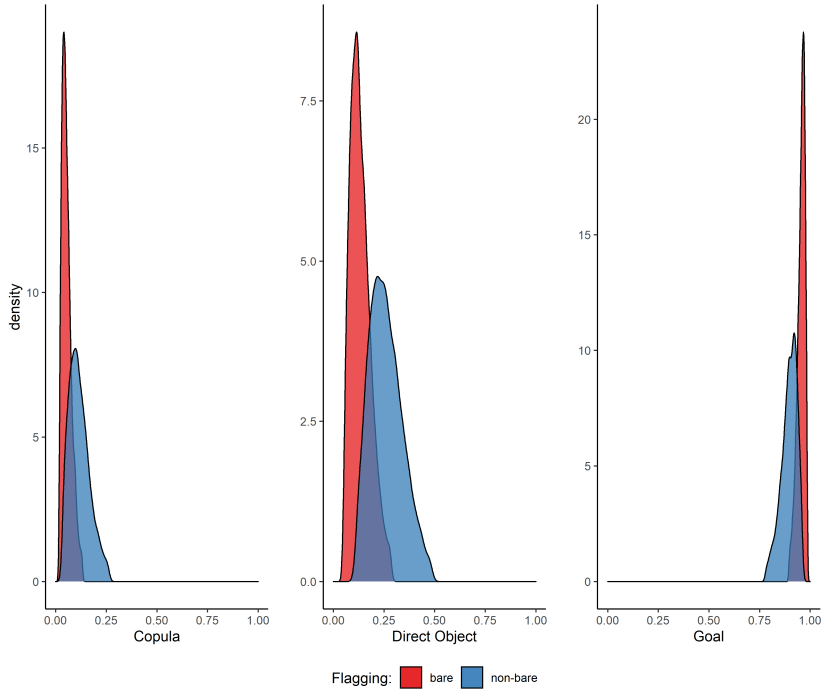


Figure 5: 95% confidence interval of parameter values for bare and non-bare roles.



in fact, the recordings were done in different villages, which could explain the large variability found in Romeyka. Overall, there are 8 texts that show effects larger than ±0.5. Out of these 8 texts, 4 texts are found in Romeyka's sub-corpus. The other 4 texts with effects larger than ±0.5 are represented by the following doculects: Turkmen variety of Balochi, Turkish of Ankara, Mazandarani of Kordxeyl, and Jewish Neo-Aramaic from Dohok.

Heavy-NP shift is considered an important predictor of the position of a constituent (Arnold et al., 2000). Thus, we would expect β_{weight} parameter to have a strong effect on the position of a constituent. This appears to be false given model's estimates of the parameters. The CI of log odds for weight parameter is [-0.15, 0.3]. The interval is reasonably narrow and close to 0. The full distribution is shown in Figure 7. One should keep in mind that weight parameter is treated here as a parameter of a continuous variable. This is not entirely correct, as weight value '4' includes data points with 4 or more words but it is safe to ignore in this case, as there are very few constituents that have more than 4 words. Thus, if anything, weight is more likely to have a very marginal effect on the position and it makes a constituent more likely to be pre-verbal. But since the effect is so low and close to 0, an explanation that weight has no effect on the position is to be preferred, at least speaking of languages in the sample. To make a more decisive generalization, one would have to assemble a larger and more diverse corpus sample of languages to test it. Overall, heavy-NP shift certainly does not have a categorical positive effect on the probability of shifting a heavy constituent to the right of the clause but it would be possible that it exists as a tendency more globally.

Figure 6: Mean parameter values of 195 different texts.

(a) Values far from 0 mean that given everything else equal, individual choice had some impact on decision about post- or pre-verbal position of a constituent.

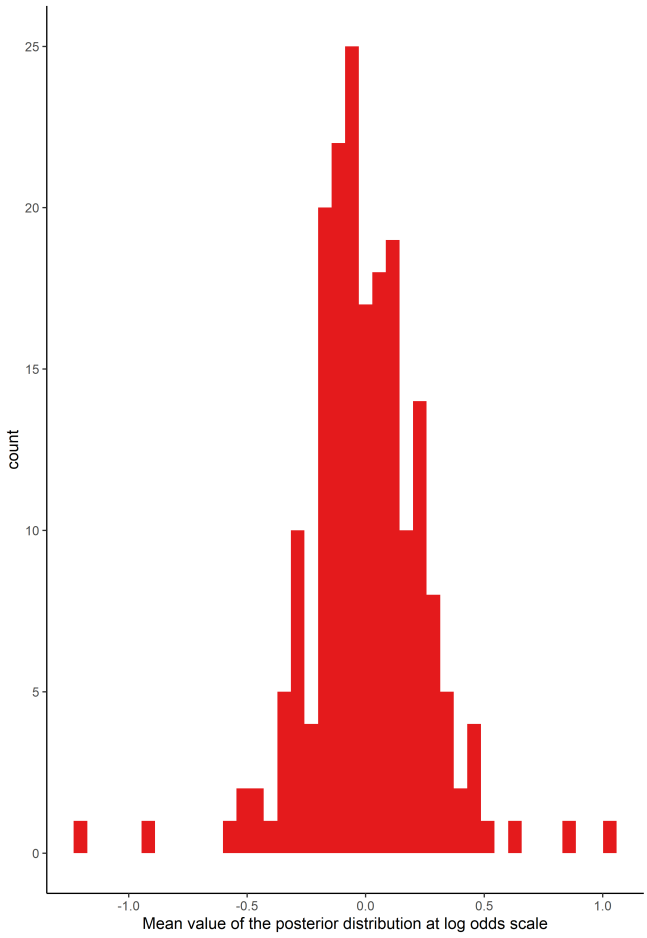
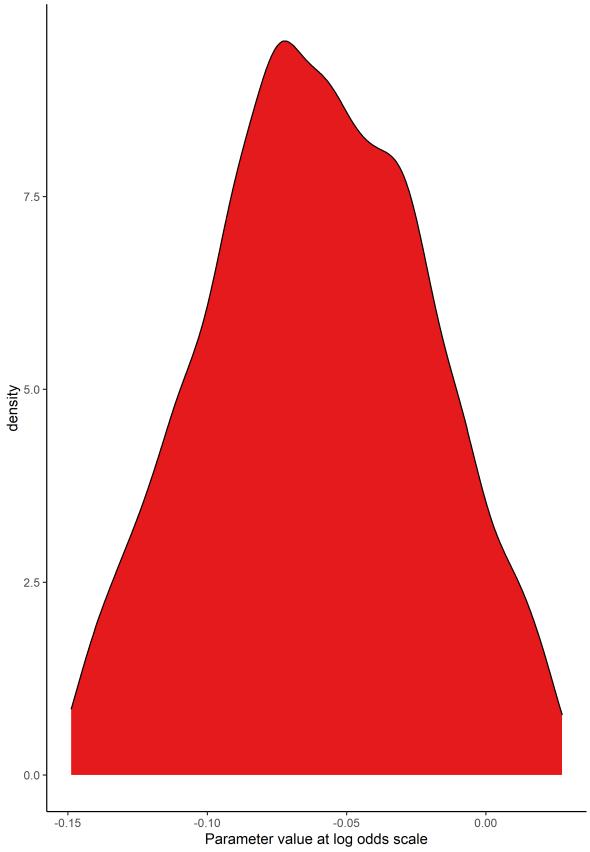


Figure 7: 95% confidence interval of β_{weight} parameter values.



In Figure 8, map with plotted doculects is shown. The lines between doculects show how correlated a given pair of doculects is, according to the model.[19] The lines opacity are scaled in accordance with the correlation strength. Note that Northern Kurdish of Ankara is slightly shifted to the north-east from Ankara to avoid complete overlap with Turkish variety of Ankara. The relationship between geographical distance and correlation strength is not linear, meaning that as the distance increases, the correlation

[19]In Figures 8 and 9 I used median values of the parameters used in the kernel and inferred by the model instead of the whole posterior distribution.

strength decays more rapidly. This can be seen in Figure 9: when geographical distance is larger than 1500 km, there is virtually no correlation.

The relationships shown in Figure 8 demonstrate that north-western Iran and Eastern Turkey is the zone with the largest degrees of correlation strength, which comes as no surprise given that this is the best represented area in the dataset. As discussed in the following section, the lines should be interpreted carefully, as the model cannot account for asymmetry in language contact. Nonetheless, an interesting observation can be made based on the values shown in Figures 8 and 9: Maximal correlation in terms of areal contact between a pair of doculects is 0.68, both of these doculects spoken in Ankara. At the same time, maximal correlation in terms of phylogenetic inheritance is 0.96. That means that model believes that despite strong areal influence in the region, ancestry of languages still has more influence on word order.[20] Another important observation related to the previous one is that Kumzari is not that strongly correlated in terms of contact intensity with other doculects in the dataset. Given that Kumzari is considered mixed Iranian-Arabic language (Wal Anonby, 2015), it could be expected. At the same time, model could fail to predict the strong areal relationship because there are no variants of Arabic spoken on the Arabian peninsula in WOWA. When Figures 10 and 11 will be introduced, it would be easy to see that in fact, some of these estimates are not entirely correct, one has to keep in mind the constant problem of undersampling of languages, both in terms of contact and phylogeny.

As I have previously said, the Gaussian processes of phylogeny and con-

[20]Correlation coefficient would not alter because of variables' scaling, by definition. Hence, the comparison here is a valid one.

Figure 8: Correlation strength between doculects.

(a) More intense contact between a pair of doculects is marked with more opaque lines.

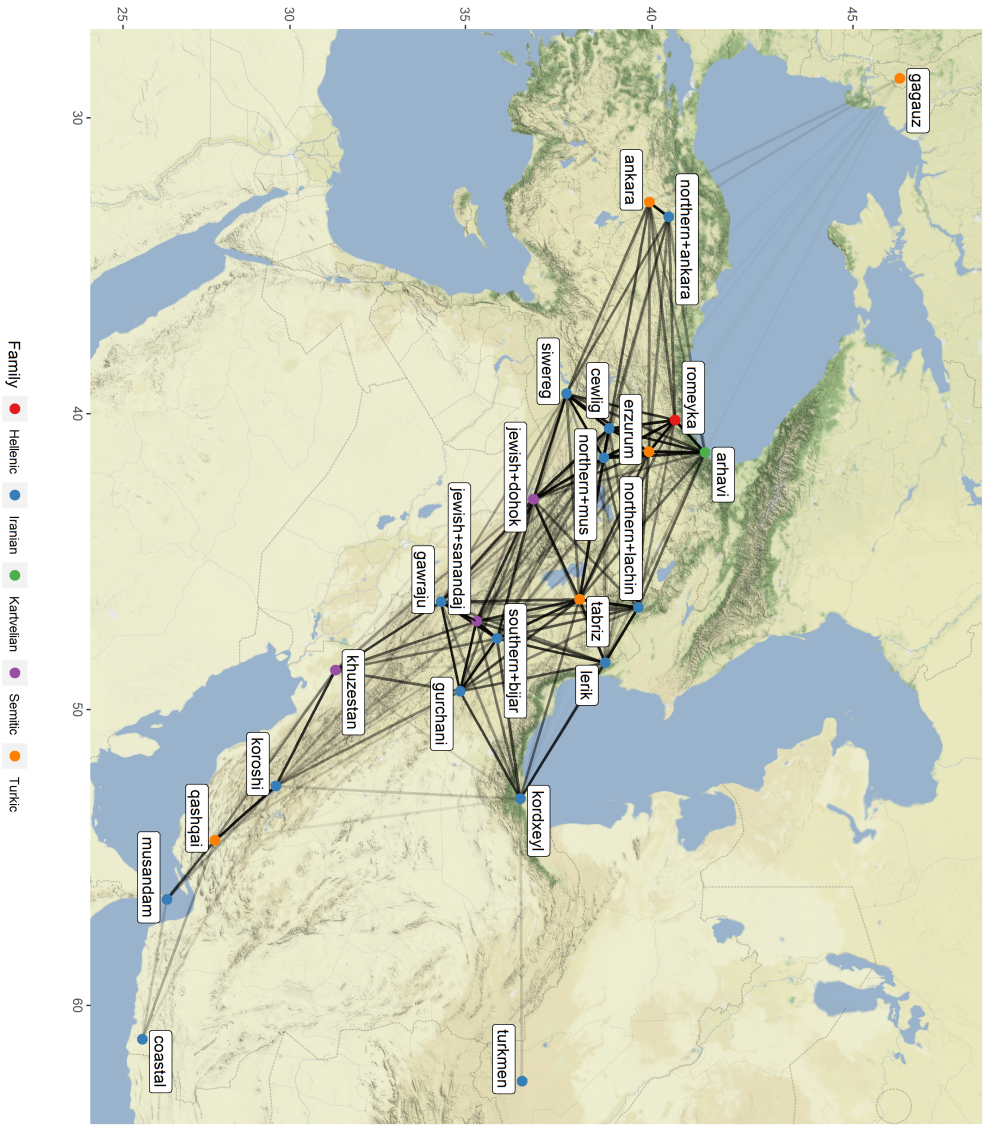
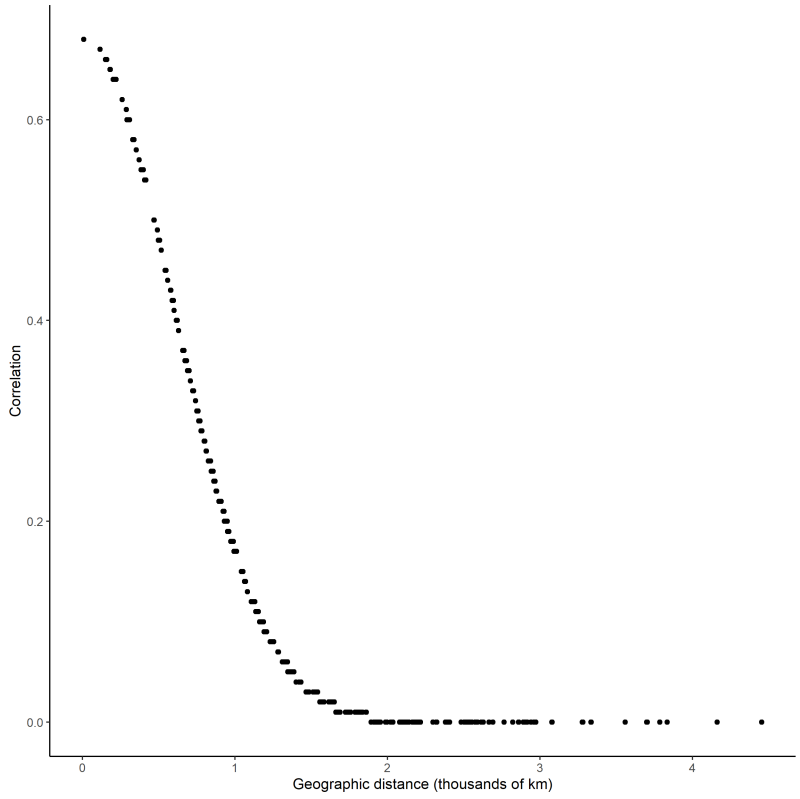


Figure 9: Relationship between correlation strength and geographical distance.

(a) Geographical distance shown in thousands of km.



tact influence in language's positioning of constituents produce intercepts. They are different from classical notion of intercepts in that they are correlated between them. In Figure 10, I show the posterior distribution of phylogenetic and contact intercepts for all doculects used in the model. It is easy to observe that the majority of the doculects' parameters have a lot of uncertainty in the estimates, especially when it comes to contact intercept. As the Figure 10 displays the parameter values on probability scale, the interpretation should go as following: the closer the value of the parameter is to 0, the more pre-verbal a doculect is; conversely, if value is closer to 1, a doculect is said to be mostly post-verbal; finally, if distribution is centred around 0, dominant word order is uncertain. Some of the rare doculects that have relatively narrow phylogenetic intercept are Oghuz varieties (apart from Gagauz), Semitic languages and Laz. To understand better the uncertainty in the estimates, I will report the average width of confidence intervals. I took 95% CI for each phylogenetic intercept and extracted minimal and maximal value found in the interval. Then I subtracted the minimal value from the maximal one and that corresponds to width of intervals. Thus, average CI parameter values range of phylogenetic intercepts is 0.60. So, even if doculect's phylogenetic intercept value has its mode around the extreme values of 0 or 1, CI will often include a lot of values that are closer to the center of x-axis (closer to the value of 0.5), i.e. closer to having no dominant word order. The average width of CIs for contact intercepts is even higher and equal to 0.66. This is certainly better than the uniform prior distribution but is somewhat unsatisfying. To reduce these interval ranges, one would have to either make

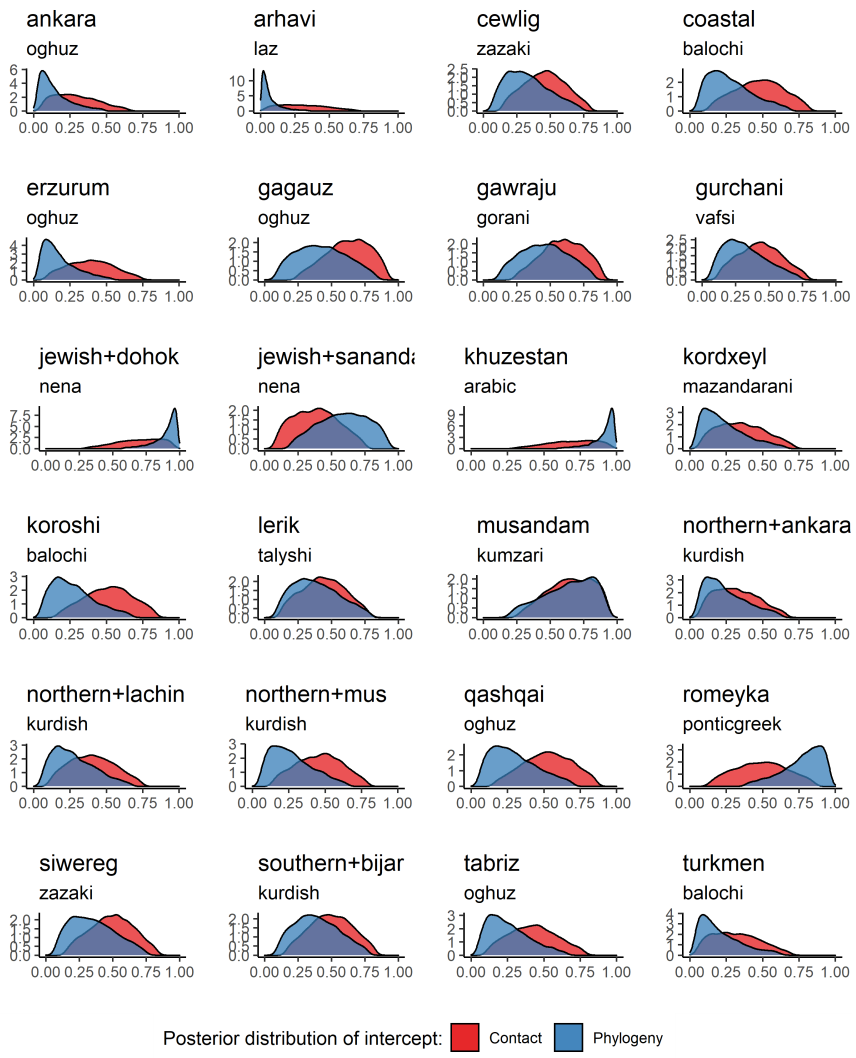
sample more diverse and include key languages that are missing[21] or provide better prior distributions. The latter solution is certainly more elegant but requires stronger theoretical assumptions. The results were left as they are, as Bayesian models conveniently take the uncertainty into account in their predictions.

One of the theoretical questions posed in first sections of this work is about historical contingency and how word order is affected by particular developments in individual languages' genetic and contact history and how that, in turn, shapes the modern distribution of word order features. Figure 10 shows model's inferences about particular doculect's phylogenetic effect and contact effect in shaping its dominant word order. But ultimately, what conditions word order in each doculect is the combination of phylogenetic and contact history, which would be unique to each doculect. For example, in Figure 10, phylogenetic intercept for varieties of Northern Kurdish are almost identical, while there are some differences in posterior distribution of contact intercepts. Another good example of that are Balochi varieties.

Thus, to make sense numerically what these unique combinations of phylogenetic and contact intercepts mean for dominant word order in a doculect, we could just sum them up. To do that, I extracted posterior distributions for phylogenetic and contact intercepts, summed them, and transformed the values to probability scale. Then, I extracted 95% CI from this distribution for each doculect. Adding these two intercepts conceptually means that we are looking at what word order distribution is predicted for a doculect given its

[21]The key languages that are missing are discussed elsewhere in the paper but generally they include: East Iranian languages, Slavic languages geographically spoken near Gagauz, Arabian peninsula's varieties of Arabic, Persian.

Figure 10: 95% confidence interval of phylogenetic and contact intercepts for each doculect.



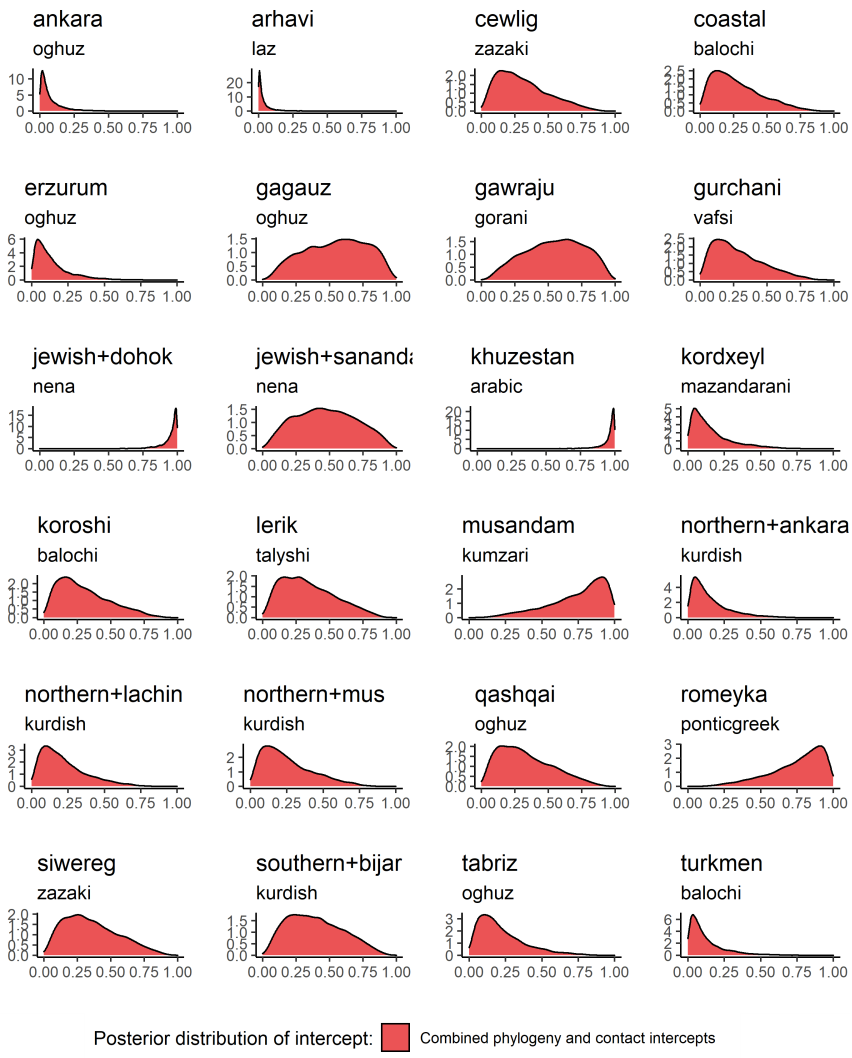
phylogenetic and contact situations, while setting all other variables to some constant value. The 95% CI for these combined distributions of phylogenetic and contact intercepts are shown in Figure 11.

4.3 Discussion

An important aspect of the Figure 4 is the position of the elements. One can perceive these values as tendencies in the area that were inferred from the data after fixing all the other variables. Thus, we can additionally confirm Haig's (2017) observation that goals tend to be post-verbal in the languages of Western Asia: bare goals have CI [0.89, 0.98], flagged goals' CI is [0.78, 0.96]. Conversely, copula elements tend to be pre-verbal, especially bare copula arguments. CI for bare copula arguments is [0.02, 0.13]. Interesting difference is observed in direct objects. Overall, their CIs differ a lot when comparing bare and flagged direct objects. Flagged direct objects have a CI of [0.11, 0.49], while bare direct objects' CI is [0.05, 0.29]. It is difficult to think about what some of these values actually mean. For instance, Turkish of Ankara has almost exclusively pre-verbal elements (540 pre-verbal elements and 42 post-verbal). But as the Figure 4 shows, goals, in general, are consistently post-verbal. These two facts have to be combined in the following way: Turkish of Ankara is a mostly pre-verbal language but if there are elements that are going to be post-verbal, these elements are likely to be goals. If we inspect the type of roles that are post-verbal in Turkish of Ankara, this observation is confirmed: out of 42 post-verbal constituents, 11 represent goals, 28 belong to the category 'other', and 3 are direct objects.

Figures 10 and 11, despite high degree of uncertainty, are worth inspecting

Figure 11: 95% confidence interval of combined phylogenetic and contact intercepts for each doculect.



in greater detail and check how well do the inferred parameter values match reality. This will not be done in this paper, as it requires language and family-specific expertise. The way the plot should be analysed goes as follows: given that many distributions have long tails, the best strategy is to look at distributions' mode, i.e. the place where the distribution has its peak. This estimate represents the degree to which a doculect has its ancestor being pre-verbal or post-verbal. When distribution peaks closer to 0, that means the model thinks that the contribution of doculect's ancestry makes it more pre-verbal, while the opposite is true about values closer to 1. Same applies to contact influence, although in the plot one cannot inspect which languages influenced the contact estimates. They can nevertheless be deduced given knowledge about doculect's history. For example, in case of Romeyka, the model believes that Romeyka's phylogeny makes it more likely to be post-verbal, while contact influence makes Romeyka somewhat more pre-verbal language. As described in Schreiber (2018), this is true to large extent, as Romeyka was mostly influenced by Turkish. As mentioned at the beginning of the paragraph, the results show a lot of uncertainty in the estimates and should be taken with caution.

Before introducing model's predictions, it was decided to quantify the performance with a metric designed as a cost function in machine learning for logistic regression. The metric is called log-loss and I took it from Géron (2019, p. 144). For our goals, this metric serves as a good heuristics about the predictions that are especially interesting to look at. In machine learning, one has to quantify model's performance numerically, so that model knows how to compare different parameter values and the fit they provide. So,

in case of machine learning, log-loss function is used directly while fitting the model, whereas in this case, it is useful to look at log-loss score and compare it to log-loss baseline score. Thus, the metric is also useful as a point of reference to understand the predictive quality of the model. Both log-loss score and log-loss baseline are explained in details in Appendix 2. The main goal that model strives to achieve is to minimize log-loss score. Thus, in Figures 13 and 12, whenever log-loss is lower than baseline, this is sign that model predicts better than it would if its prediction always was the more frequent position per doculect. As we can see out of Figure 13, overall model's predictions for the entire dataset are way better than log-loss baseline score. But Figure 12 is more informative, as it computes log-loss score by doculect. In this case, we can see that the majority of doculects have log-loss score lower than baseline. It is interesting to pay more attention to the two doculects that have baseline log-loss score lower than baseline log-loss. These are: Laz of Arhavi and Arabic of Khuzestan. To understand why log-loss score for Laz and Arabic are higher than the baseline, one has to keep in mind two factors: (1) These doculects are highly homogeneous in their word order. More than 95% of Laz constituents are pre-verbal, while more than 90% of Arabic constituents are post-verbal. (2) The model lacks any doculect-specific parameters. The first factor explains why these two have some of the lowest baselines compared to other doculects, meaning that model has to make a lot of accurate predictions to get lower than the baseline. Homogeneity is associated with lower baseline for reasons explained in Appendix 2. Coming to the second factor mentioned above, a model with doculect-specific parameter could easily assign most of the effect's strength

to this single predictor and this way it would make accurate predictions in case of more homogeneous doculects. Simply put, if model was allowed to say that Laz is 95% of the times pre-verbal independent of other factors, it would still be accurate in more than 95% of the cases. But such model carries little predictive potential, as it requires doculect-specific parameter and is uninteresting from theoretical point of view.

Some other doculects that model predicts only marginally better than it would under baseline conditions are: Turkish of Ankara, Turkmen Balochi and Gagauz. The reason why model predicts only marginally better the positions of Turkish of Ankara and of Turkmen Balochi are the same as in case of Arabic and Laz. These languages are highly homogeneous in their word order, both have more than 90% of the constituents in the dataset being pre-verbal. The case of Gagauz is more interesting and surprisingly the model still predicts its constituents positions better than the baseline, although the difference is so low that it can be ignored. Unfortunately, WOWA does not have some of the Indo-European languages (Romanian, Bulgarian, Russian) necessary for Gagauz's predictions to be accurate. Hence, as will be shown, Gagauz is unsurprisingly one of the worst doculects in terms of predictions.

As I mentioned in the previous section, the dataset was divided into two parts: training part and testing part. The results I reported above are based solely on the training part of the dataset. To test how accurate the parameter estimates are, one has to plot the predictions of the model to appreciate whether the model has any predictive power. This is especially true in the case of logistic regression models. Observe the function 'logit' in Equation 5. Applying the logit transformation, changes the range of possible

Figure 12: Log-loss score by individual doculects.

(a) Under good model, log-loss of the model (blue) is expected to be lower than baseline log-loss (red).

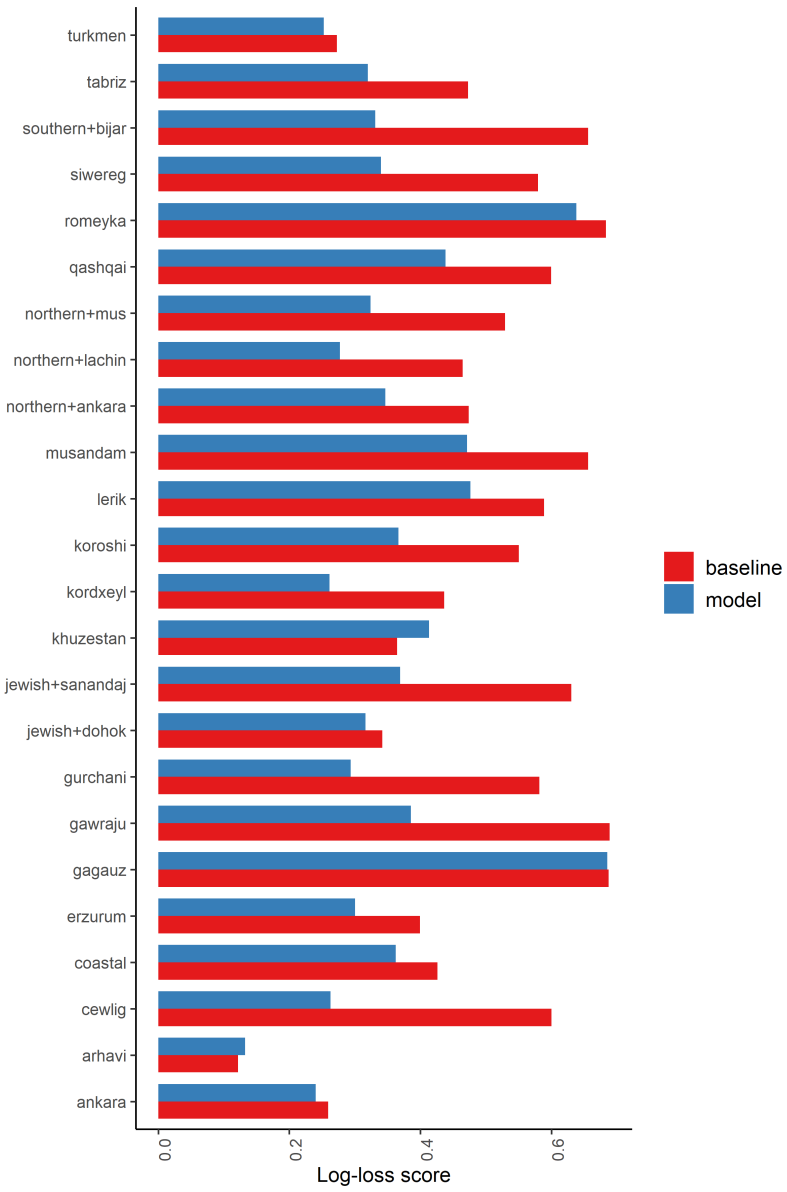
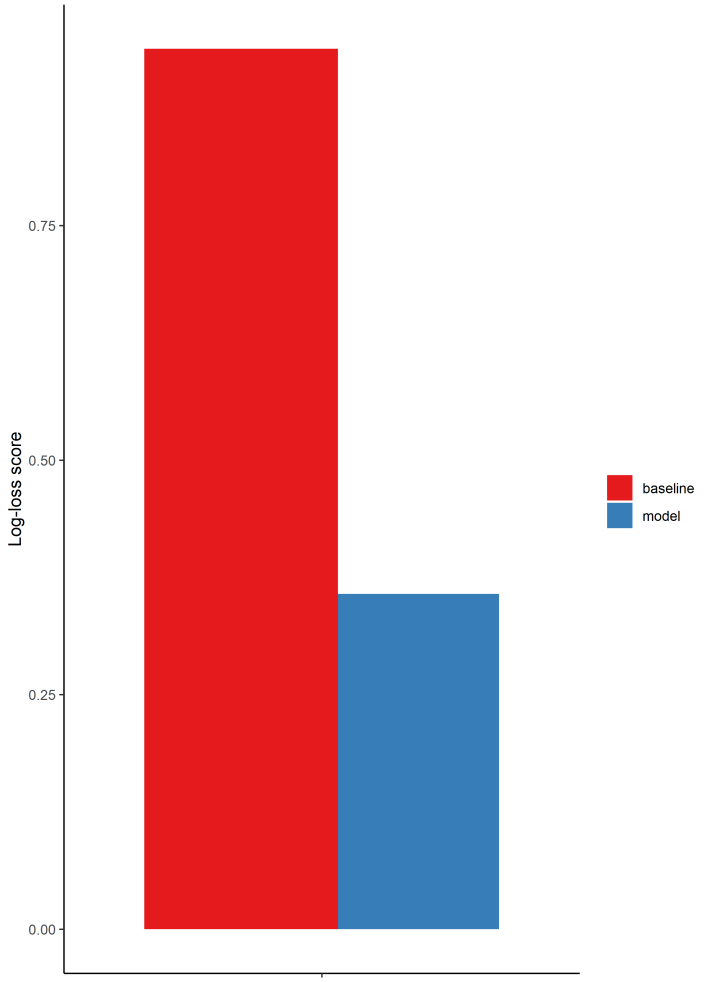


Figure 13: Log-loss over the whole dataset.

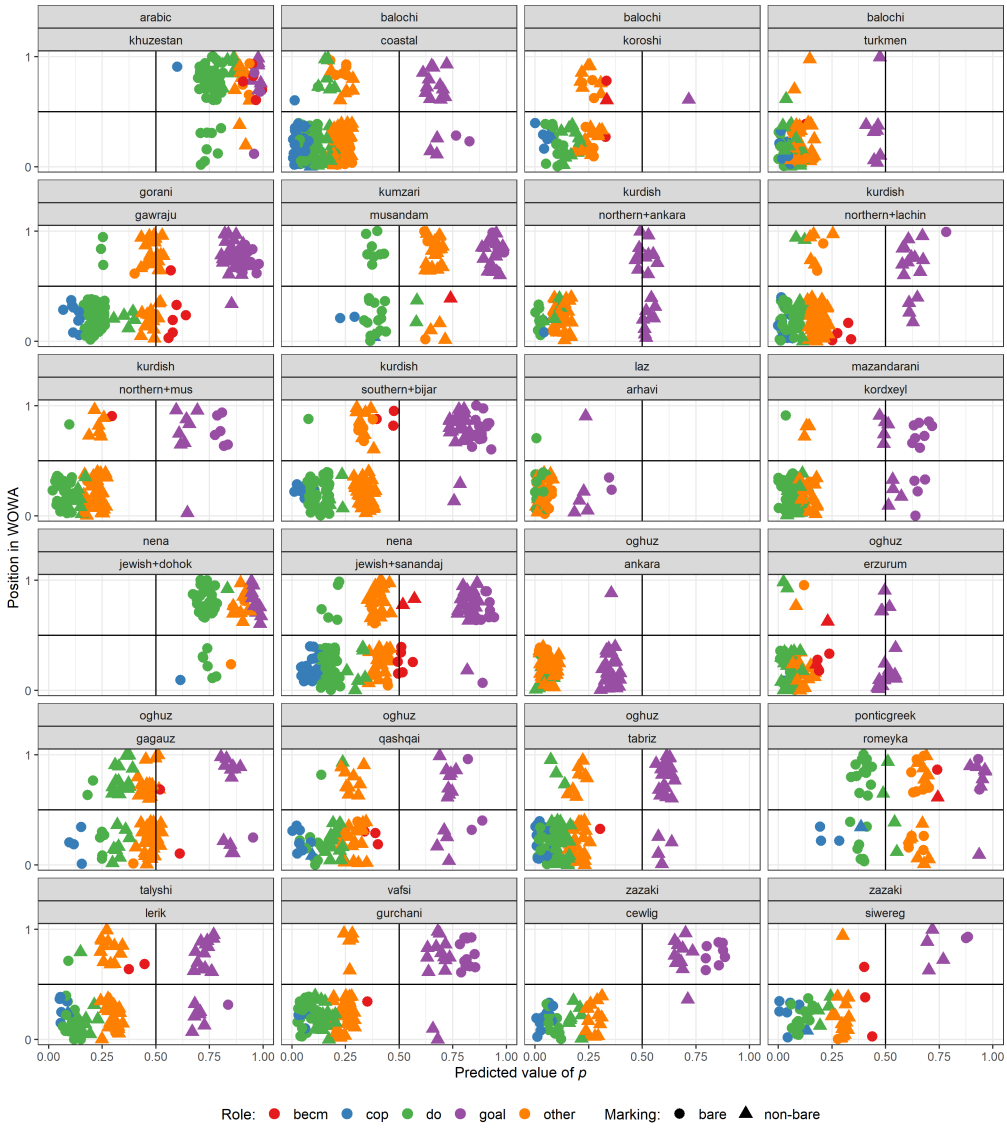
(a) Under good model, log-loss of the model (blue) is expected to be lower than baseline log-loss (red).



values, shifting it from the initial [-∞, +∞] to [0, 1]. By doing this, we make the parameters covary too, even though they would not necessarily covary on the initial scale (Duncan & Kefford, 2021, p. 2289). They covary because the maximal and minimal values are now constrained, and hence an increase in one value cannot go on increasing infinitely, the effect of some other variable will have to decrease in order to allow it. Because of that, it is important to also check the predictions instead of merely looking at parameter values. This allows to account for the totality of the variables and their interactions, intended and unintended ones by design, instead of looking at possibly deceiving parameter values.

The predictions are displayed in Figure 14. The plot is complex and requires some explanation. On the Y-axis one sees "real" position, i.e. the position that an element has in WOWA dataset, as coded by an annotator. Hence, all the observations should in fact lie along the lines '0' or '1'. I have slightly shifted the observations along the Y-axis to make it possible to understand the rough number of elements corresponding to a certain point along the X-axis, as they would otherwise overlap. X-axis shows the predicted value of the parameter p. In simple terms, whenever the predicted value of p is small, we would expect the position of an element to be pre-verbal (=0), and the opposite is true. Having this in mind, it should be clear which predictions are more consistent with the real data. I have separated each sub-plot into four squares. The predictions consistent with coded data should either be in the bottom left corner or in the top right corner of each sub-plot. At the same time, whenever points are close to the central vertical line, it means that the value of p is close to 0.5, which corresponds to almost random

Figure 14: Predictions on the test sample.



placement of an element. For instance, note that all the non-bare goals of Oghuz variety of Erzurum are placed exactly this way. The situation is similar in case of Northern Kurdish variety of Ankara. In both cases roughly half of the points lie above the horizontal line and half below, meaning that there is a lot of uncertainty in the positioning of an element, and the model considers that their position is almost randomly drawn.[22]

The first things to note about Figure 14 is that not all doculects have the same degree of certainty about all roles. Most often, the category 'other' is closer to the $p \approx 0.5$. This is unsurprising given that 'other' contains several categories. It is worth now taking a closer look at some of the subplots, as this will show the advantages and disadvantages of the model. As McElreath (2020) remarks, it is also a sign of a good model, when it fails in transparent ways. I believe that to a large extent, this is the case.

To observe how the model accommodates contact, we have to look at the doculects that are genetically closely related but more or less distributed in space. Balochi is the first clade, alphabetically and in Figure 14, that fits this criteria. The corpora for all Balochi varieties in the dataset are well represented and the model nicely captures the contact situation. Despite some mispredictions, the vast majority of the constituents' positions are predicted correctly, and many of the mispredictions occur with the category 'other' and with 'goals'. As previously mentioned, the model parameters show that non-bare goals have more variability in their position. Almost all the mispredicted goals in case of Balochi are flagged in some way. Moreover, in case

[22]We could take a deterministic perspective here and say that nothing is random. For instance, if WOWA was coded at discourse level with some pragmatic details, it would be likely that the model would be able to account for this "randomness" and make more certain predictions.

of Coastal Balochi, most of the mispredicted elements in general are flagged. Despite the lack of Eastern Iranian languages and Turkic varieties to the East of Iran, the model still makes decent predictions for Turkmen variety of Balochi.

The next clade that is worth looking at is Kurdish, particularly Northern Kurdish varieties spoken in Ankara, Lachin and Muṣ. We can see that most of the elements in Ankara Kurdish are pre-verbal. The only mispredicted roles are goals and 'other', the latter being mispredicted only when it is flagged. Otherwise, the model predicts goals to be post-verbal with close to 50% probability, while all bare goals in all the varieties of Kurdish are post-verbal. Given that Ankara Kurdish is highly influenced by the standard Turkish, it is to be expected for it to be mostly a pre-verbal language, which is reflected by model's predictions of having all the elements having value of p closer to 0. Overall, across the whole Kurdish clade, the most often mispredicted roles are 'other' and non-bare goals. This pattern can be seen not only in Northern varieties but also in Southern Kurdish variety of Bijar.

The next in our list are two Jewish Neo-Aramaic varieties. Despite these two varieties being spoken relatively close to one another, the important difference here is made by political map. The variety of Dohok is spoken in Iraq, where the official and major language is Arabic, while Sanandaj's most widely spoken languages are Kurdish varieties and the official language is Persian. These two West Iranian languages have mostly pre-verbal placement of many roles, which has apparently had a lot of influence on word order in Jewish variety of Neo-Aramaic spoken in Sanandaj. The contact influence in this particular case is probably most obvious, as most of the elements of

Dohok's variety are post-verbal, while most of the elements apart from goals are pre-verbal in Sanandaj's variety. The category 'other' is again most often mispredicted for the same reasons, as described above.

Oghuz branch of Turkic is the best represented clade in WOWA, it has 5 varieties that are spread across the region. The overall pattern that is applicable to all the varieties except Gagauz is the following: the more south-eastern a variety is, the more post-verbal goals it has. Ankara Turkish, also considered standard Turkish, consistently has most of the elements in pre-verbal position. The next Oghuz variety to the East is the variety of Erzurum. As it is spoken in Turkey, it has experienced the influence of standard Turkish and has again most of it elements being pre-verbal, although p value for parameters is again close to 0.5 meaning that speakers probably have a lot of freedom with regard to goals' placement. Qashqai and Tabriz varieties tend to place their goals post-verbally, although in both of these, there are a few instances of pre-verbal goals. The rest of the roles are mostly pre-verbal, as predicted by the model, although there are a few instances of 'other' and a few direct objects that are post-verbal but overall their proportion is low enough to call them outliers. Gagauz will be discussed in more details when I discuss disadvantages.

Finally, we have Zazaki (Dimli) varieties. But the two varieties represented here are spoken next to one another in Eastern Turkey. Because of that, a similar pattern is observed: goals and some elements of the category 'other' are post-verbal, whereas the rest of the elements tend to be pre-verbal. The variety of Çewlîg has one non-bare goal that is pre-verbal. In Siwêreg variety, pre-verbal goals might remain unobserved due to sampling which oc-

curred when the dataset was separated into training and testing part. After inspecting the data manually, this guess is confirmed. Despite there being only few pre-verbal goals in Siwêreg variety, they can still be found, as well as some more pre-verbal goals in Çewlîg.

Turning now to the poor predictions, there are a few patterns we observe here. The first, often mentioned above problem is the prediction of the category 'other' which comes as no surprise. Some of the roles that were of less interest had to merged, as otherwise the model would have too many parameters to infer.[23] This comes at the price of poor predictions for this category, apart from a few doculects that are consistent in their placement of constituents, e.g. Laz and Turkish of Ankara.

The other problem related mostly to the dataset is absence of some doculects. There is no guarantee that the model would not fail in some of the examples discussed above, but I expect it to make better predictions in case the model would have access to some other languages relevant to the contact situation. This mostly refers to the languages spoken at the border of the region. For example, Russian, Bulgarian and Romanian, the languages with which Gagauz had a lot of contact (Pokrovskaja, 1978), are not present in the dataset. Kumzari was in contact with the Arabic of Arabian peninsula for centuries (Wal Anonby, 2015). Despite having one variety of Arabic in the dataset, this variety is spoken in Khuzestan which is far more distant variety geographically from Kumzari than the peninsular Arabic varieties. The problems with Romeyka are twofold: (1) As mentioned above, there is a lot of variability in word order of Romeyka between individuals (2) There

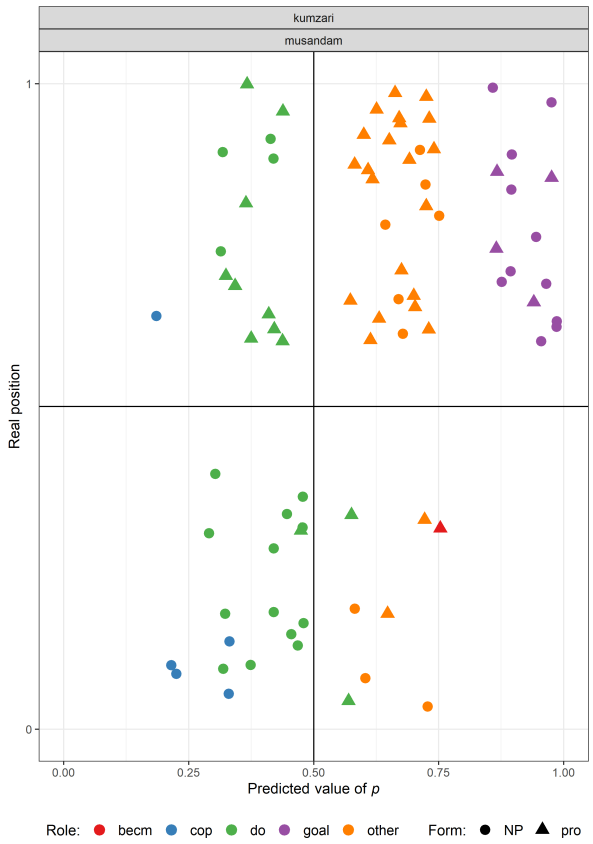
[23]Every additional category is a parameter for the model and hence we would need to have a lot of observations per category.

is no other variety of Greek in the dataset for model to have an idea of how much of Romeyka's word order is explained by contact and how much by genetic inheritance. Overall, the dataset will never be complete. As geographical coverage increases, so does the number of doculects that need to be included. This further supports the idea of fluidity in contact influence. Nonetheless, it is worth working on at least some of the languages I mentioned. Some of them can be seen as a base, of what a language would look like when it is spoken outside the region as opposed to inside the transition zone.

A more sophisticated problem that would require a better formulated model is found in Kumzari variety of Musandam. Kumzari places direct objects before the verb when they are expressed with an NP, while a direct object expressed with a pronoun will follow the verb (Wal Anonby, 2015, p. 67). Kumzari is widely accepted to be a part of Western Iranian clade and hence the placement of NP before the verb may merely be a genetic retention. Interestingly, the pattern found in Kumzari is at odds with the universal 25 (Greenberg, 1963): The universal 25 states that if pronominal object follows the verb, so does the nominal object. It is likely that pronominal objects follow the verb because of the influence of Arabic on Kumzari. Thus, we see that contact in this case is not acting on the role level, but rather it is more selective and had influence at the level of object's form. After observing the pattern, I have tried including an additional categorical variable of reduced versus nominal form and its interaction with a particular doculect. The MCMC chains do not converge in this case and that makes the interpretation of the posterior distribution meaningless. Including the reduced versus non-

Figure 15: Prediction on the test sample, only Kumzari of Musandam.

(a) The shape of a point now shows element being expressed by a Noun Phrase (NP) versus some reduced form (pro).



reduced form without interaction also does not help, as the model infers the universal pattern, rather than doculect specific one. I show predictions for Kumzari in Figure 15 to demonstrate that most of the mispredictions of direct objects happen when object is expressed in a reduced form.

There is one more technical issue to which I do not have a solution at the moment of writing this paper. Gaussian process models make observations covary and that means that influence is symmetric. Linguists are well aware of the fact that languages influence one another more or less depending on socio-political status of language and the mutual perceptions of the communities that enter in contact. Thus, in many cases language contact is not symmetric and one language acquires more from the other. An example of that is the case of above mentioned Kumzari. Even though Arabic has undergone a lot of change in contact with West Iranian languages, to my knowledge Kumzari has never been reported to have any major impact on Arabic, due to its lack of prestige status in places where it is spoken.[24]

5 Conclusion

As the study shows, given a few structural features, genetic information about a doculect and its geographical position, the model is predicting well the relative position (pre-verbal or post-verbal) of non-subject elements in languages of Western Asia that are included in WOWA. This partially supports the idea of historical contingency that states that some of the observed distributions of typological features are historical accidents formed by expan-

[24]This appreciation is based on my personal perception of sociolinguistic situation of Kumzari speakers described in Wal Anonby (2015).

sion of some language families, contact. There remains however one question related to the study by Hawkins (2008) which I have introduced in Section 2.3. It would be a strong claim to say that well-observed asymmetry in the position of obliques is merely a results of historical coincidence and one should probably be still cautious about rushing conclusion and extending historical contingency hypothesis to being able to explain the asymmetry shown in Table 1. Nonetheless, the model shows that contact and genetic history of a language are shaping the distribution of word order features in languages of the world to the extent that it is possible to make good predictions about word order in case we have this information.

I have shown that some of the typological tendencies of Western Asia as a linguistic area presented in Haig (2017) are confirmed by this study. Most notably, Haig's idea about the final position of goals as being a typological tendency of the area is confirmed by model's parameter values. A similar study in corpus-based framework on a macro-scale could provide explanation for lack of diversity of oblique's word order in VO languages. Any conclusion about it based on WOWA dataset would be too ambitious. What could be claimed based on the present study is that Western Asia's order of obliques could be explained by historical circumstances of the languages spoken in this area.

The model fails when there is a more selective contact pressure in a particular type of constituents, or across a particular linguistic feature. The limitations of the model are also explained by the set of languages that were not included in the study. The word order of some doculects is best explained by contact with languages that are not found in WOWA and hence the model

lacks relevant information to make good predictions for word order in those doculects. We also find partial confirmation of the idea that the presence of flagging enables speakers to place a constituent in different positions but this effect also depends on the grammatical role of a constituent.

The paper proposes a few directions for methodological development that is needed in case linguistic typology pursues the path of directly dealing with non-independent observations, rather than applying sampling to avoid it. Apart from that, the work further emphasizes the need for corpus-based typology and introduces a different way in which corpora may be applied to study (areal) typological questions.

There still remain a lot of aspects in methodology that could further improve the results. One particularly interesting way of improving the study is thinking about the way in which we could model the influence of contact and genetic inheritance given a particular feature of an element, like in the case of Kumzari described above.

An easier problem but also an important one to study is the influence of phylogeny scaling of Glottolog's trees and which one would be (more or less) universally appropriate. For some language families, like Indo-European or Austronesian, there are reference trees based on lexical data that could be used but typologists often deal with languages for which Glottolog's judgements represent the best available topology. Therefore, there might be a demand for having a way of scaling Glottolog's trees appropriately in a heuristic way to use the trees in an adequate manner.

I have shown that Gaussian process represents an appropriate statistical technique, at least for the near future. Potential improvement in model's fit

might be achieved by implementing phylogenetic logistic regression (Ives & Garland, 2014). Phylogenetic logistic regression is a relatively new development and, as reported in Ives and Garland (2014), might be harder to fit but conceptually would represent an even better way for making inferences.

WORD COUNT

15200

References

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

Bickel, Balthasar, Grenoble, Lenore A., Peterson, David A., & Timberlake, Alan (Eds.). (2013). *Language typology and historical contingency.* John Benjamins. doi: 10.1075/tsl.104

Cinelli, Carlos, Forney, Andrew, & Pearl, Judea. (2020). A crash course in good and bad controls. *SSRN*. (preprint) doi: 10.2139/ssrn.3689437

Collins, Jeremy. (2018). Some language universals are historical accidents. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis, & Ilja Seržant (Eds.), *Explanation in typology* (pp. 47–61). Berlin: Language Science Press. doi: 10.5281/zenodo.2583788

Dryer, Matthew S. (1992). The Greenbergian word order correlations. *Language*, *68*, 138–181.

Dryer, Matthew S. (2013a). Order of adposition and noun phrase. In Matthew S. Dryer & Martin Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/85`

Dryer, Matthew S. (2013b). Order of object and verb. In Matthew S. Dryer & Martin Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/83`

Dryer, Matthew S. (2013c). Order of subject, object and verb. In

Matthew S. Dryer & Martin Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/81`

Dryer, Matthew S., & Gensler, Orin D. (2013). Order of object, oblique, and verb. In Matthew S. Dryer & Martin Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/84`

Duncan, Richard P., & Kefford, Ben J. (2021). Interactions in statistical models: Three things to know. *Methods in Ecology and Evolution*, *12*(12), 2287-2297. doi: 10.1111/2041-210X.13714

Dunn, Michael, Greenhill, Simon J., Levinson, Stephen C., & Gray, Russell David. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*, 79–82. doi: 10.1038/nature09923

Futrell, Richard, Mahowald, Kyle, & Gibson, Edward. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341. doi: 10.1073/pnas.1502134112

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki, & Rubin, Donald B. (2013). *Bayesian data analysis* (3rd ed.). New York: Chapman and Hall/CRC.

Gerdes, Kim, Kahane, Sylvain, & Chen, Xinying. (2021). Typometrics from implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, *6*, 1–31. doi: 10.5334/gjgl.764

Greenberg, Joseph H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (Ed.), *Universals of human language* (pp. 73–113). Cambridge, Massachusetts: MIT Press.

Géron, Aurélien. (2019). *Hands-On machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media, Inc.

Haig, Geoffrey. (2017). Western Asia: East Anatolia as a Transition Zone. In Raymond Hickey (Ed.), *The Cambridge handbook of areal linguistics* (p. 396—423). Cambridge University Press. doi: 10.1017/9781107279872.015

Haig, Geoffrey, & Khan, Geoffrey (Eds.). (2018). *The languages and linguistics of Western Asia: An areal perspective*. De Gruyter Mouton. doi: 10.1515/9783110421682

Haig, Geoffrey, & Schnell, Stefan (Eds.). (2021). *Multi-CAST: Multilingual corpus of annotated spoken texts*. Bamberg. Retrieved from `multicast.aspra.uni-bamberg.de/` (Accessed on 17/01/2022)

Haig, Geoffrey, Stilo, Donald, Doğan, Mahîr C., & Schiborr, Nils N. (Eds.). (2021). *WOWA — Word Order in Western Asia*. Bamberg: University of Bamberg. Retrieved from `multicast.aspra.uni-bamberg.de/resources/wowa/`

Hammarström, Harald, Forkel, Robert, Haspelmath, Martin, & Bank, Sebastian. (2021). *Glottolog 4.5.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved 05/04/2022, from `http://glottolog.org`

Harmon, Luke. (2020). *geiger* [Computer software manual]. Retrieved

from https://github.com/mwpennell/geiger-v2 (R package version 2.0.7)

Hawkins, John A. (2008). An asymmetry between VO and OV languages: The ordering of obliques. In Greville G. Corbett & Michael Noonan (Eds.), *Case and grammatical relations: Studies in honor of Bernard Comrie* (pp. 167–190). Amsterdam: John Benjamins. doi: 10.1075/tsl.81.08ana

Hickey, Raymond. (2017). Areas, areal features and areality. In Raymond Hickey (Ed.), *The cambridge handbook of areal linguistics* (p. 1–16). Cambridge University Press. doi: 10.1017/9781107279872.002

Huehnergard, John. (2019). Proto-Semitic. In John Huehnergard & Na'ama Pat-El (Eds.), *The Semitic languages* (2nd ed., pp. 49–79). London, New York: Routledge.

Ives, Anthony R., & Garland, Theodore. (2014). Phylogenetic regression for binary dependent variables. In László Zsolt Garamszegi (Ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (pp. 231–261). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-43550-2_9

Ivić, Pavle, & Crystal, David. (2014). *"dialect"*. Retrieved from https://www.britannica.com/topic/dialect (Accessed on 19/02/2022)

Jeszenszky, Péter, Stoeckle, Philipp, Glaser, Elvira, & Weibel, Robert. (2018). A gradient perspective on modeling interdialectal transitions. *Journal of Linguistic Geography*, *6*(2), 78—99. doi: 10.1017/jlg.2019.1

Jing, Yingqi, Widmer, Paul, & Bickel, Balthasar. (2021). Word order varia-

tion is partially constrained by syntactic complexity. *Cognitive Science*, *45*(11). doi: 10.1111/cogs.13056

Kiparsky, Paul. (1996). The shift to head-initial VP in Germanic. In Hubert Haider, Susan Olsen, & Sten Vikner (Eds.), *Studies in comparative Germanic syntax II.* Dordrecht: Kluwer Academic Publishers.

Levshina, Natalia. (2019). Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, *23*(3), 533–572. doi: 10.1515/lingty-2019-0025

Lewandowski, Daniel, Kurowicka, Dorota, & Joe, Harry. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*, 1989–2001.

Macklin-Cordes, Jayden L., & Round, Erich R. (2022). *Challenges of sampling and how phylogenetic comparative methods help: With a case study of the Pama-Nyungan laminal contrast.* (preprint) doi: 10.48550/arXiv.2201.00195

McElreath, Richard. (2020). *Statistical rethinking: A Baeysian course with examples in R and Stan* (2nd ed.). Chapman & Hall/CRC.

Naranjo, Matías Guzmán, & Becker, Laura. (2021). Statistical bias control in typology. *Linguistic Typology*. doi: 10.1515/lingty-2021-0002

Naroll, Raoul. (1961). Two solutions to Galton's Problem. *Philosophy of Science*, *28*(1), 15–39.

Neureiter, Nico, Ranacher, Peter, Efrat-Kowalsky, Nour, Kaiping, Robert, Gereon A. Weibel, Widmer, Paul, & Bouckaert, Remco R. (2022). Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications*,

9. doi: 10.1057/s41599-022-01211-7

Pagel, Mark. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*, 877–884. doi: 10.1038/44766

Pokrovskaja, L.A. (1978). *Sintaxis Gagauzskogo Yazyka v sravnitelnom osveșenii [Syntax of the Gagauz language from comparative perspective].* The Academy of Science of the USSR.

Robbeets, Martine, Bouckaert, Remco, Conte, Matthew, Savelyev, Alexander, Li, Tao, An, Deog-Im, ... Ning, Chao (2021). Triangulation supports agricultural spread of the transeurasian languages. *Nature*, *599*(7886), 616-621. doi: 10.1038/s41586-021-04108-8

Round, Erich R. (2021). glottoTrees: Phylogenetic trees in linguistics. [Computer software manual]. Retrieved from `https://github.com/erichround/glottoTrees` (R package version 0.1)

Schreiber, Laurentia. (2018). Romeyka. In Geoffrey Haig & Geoffrey Khan (Eds.), *The languages and linguistics of Western Asia: An areal perspective* (pp. 892–934). De Gruyter Mouton. doi: 10.1515/9783110421682

Stan Development Team. (2022). Stan modeling language users guide and reference manual, 2.29 [Computer software manual]. (https://mc-stan.org)

Stilo, Donald. (2005). Iranian as buffer zone between the universal typologies of Turkic and Semitic. In Éva Ágnes Csató, Bo Isaksson, & Carina Jahani (Eds.), *Linguistic convergence and areal diffusion: Case studies from Iranian, Semitic and Turkic.* Routledge.

Wal Anonby, C.A. van der. (2015). *A grammar of Kumzari: A mixed Perso-*

Arabian language of Oman (Doctoral dissertation, Leiden University). Retrieved from https://hdl.handle.net/1887/32793

Windfuhr, Gernot. (2009). Dialectology and topics. In Gernot Windfuhr (Ed.), *The Iranian languages* (1st ed.).

Zeman, Daniel, Nivre, Joakim, Abrams, Mitchell, Aepli, Noëmi, Agić, Željko, Ahrenberg, Lars, ... Zhu, Hanzhi (2019). *Universal dependencies 2.5.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Retrieved from http://hdl.handle.net/11234/1-3105

Appendices

Appendix 1

$$position_i \sim \text{Binomial}(1, p_i)$$

$$\text{logit}(p_i) = P_i + G_i + RF_{[role_i, flag_i]} + T_i + \beta_w W_i$$

$$\begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_{24} \end{pmatrix} \sim \text{MVNormal}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{K}_{phylo} \right)$$

$$\text{K}_{phylo[i,j]} = \eta^2 \exp(-\rho^2 D_{phylo[i,j]}) + \delta_{i,j}\sigma^2$$

$$\eta^2 \sim \text{Exponential}(0.35)$$

$$\rho^2 \sim \text{Exponential}(0.35)$$

$$\delta_{i,j}\sigma^2 \sim \text{Exponential}(6)$$

$$\begin{pmatrix} G_1 \\ G_2 \\ \vdots \\ G_{24} \end{pmatrix} \sim \text{MVNormal}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{K}_{geo} \right)$$

$$\text{K}_{geo[i,j]} = \eta^2 \exp(-\rho^2 D^2_{geo[i,j]}) + \delta_{i,j}\sigma^2$$

$$\eta^2 \sim \text{Exponential}(0.35)$$

$$\rho^2 \sim \text{Exponential}(0.35)$$

$$
\begin{aligned}
\delta_{i,j}\sigma^2 &\sim \text{Exponential}(2) \\
\begin{bmatrix} RF_{j,1} \\ RF_{j,2} \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \rho \quad \sigma^2_{R_i} \right) \\
\rho &\sim \text{LkjCorr}(2) \\
\sigma^2_{R_i} &\sim \text{Exponential}(1) \\
T_i &\sim \text{Normal}(0, 0.5) \\
\beta_w &\sim \text{Normal}(0, 1.5)
\end{aligned}
\tag{6}
$$

In Equation 6, I follow McElreath (2020) in notation. MVNormal stands for multivariate normal distribution. K_{phylo} and K_{geo} represent kernels of Gaussian processes. Thus, the first two multivariate normal distributions represent intercepts, which are correlated between them and the correlation is dependent on distance between each individual pair of doculects in the dataset. As discussed in the paper, D represents either phylogenetic or geographical distance between doculects. The kind of distance can be identified through subscript.

The parameters η^2, ρ^2 and δ of each Gaussian process are inferred separately. In case of RF, interaction term of different combinations of role and flag, we have a few terms to consider. The important aspect that is modelled with this third multivariate distribution is that each role will have a different effect on the outcome and its effect will change along with the type of flagging. $\sigma^2_{R_i}$ just represents the uncertainty in the estimates.

An important property of exponential distribution to keep in mind is that

it is wider, when the parameter describing the distribution is lower. Thus, δ parameter for phylogenetic kernel is much more constrained to be lower than the same parameter of geographical distance kernel (contact kernel). In case of role and flag interaction, the prior for ρ is described with the help of LkjCorr distribution (Lewandowski et al., 2009). It is needed for this particular ρ parameter because we infer a few ρ values and then assemble them into a matrix which then describes correlation between each role and flag and how they covary together.

Appendix 2

The following equation represents the log-loss cost function:

$$C(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \tag{7}$$

In the formula above: $C(\boldsymbol{\theta})$ is a cost function of a vector of model's parameters $\boldsymbol{\theta}$; n is the number of data points; y is the outcome (position as coded in WOWA); $\hat{p}$ is the estimated probability of "success" (i.e. being post-verbal). In Figure 16, I show two cases: (1) when position is fixed at 1 and predicted value of p ranges from 0.01 to 0.99. (2) when position is fixed at 0 and the same interval is used for p. When p is closer to low values and the position is fixed at 0, log-loss function yields a small number, lower than 1. The opposite is true for position being fixed at 1 and predicted p being high.

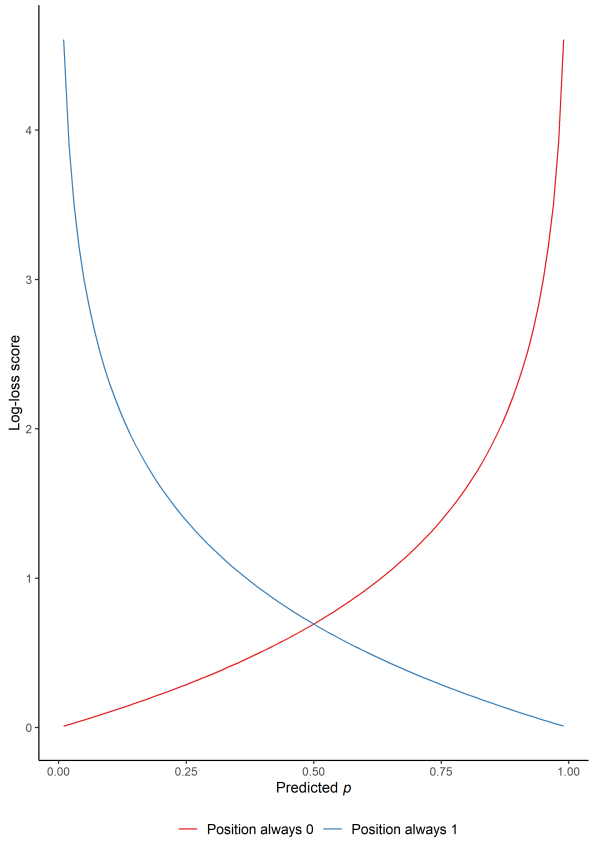
Note how function reflects the continuous character of mispredictions being made by model. For example, when the model predicts very high p, e.g. p=0.9, while the position is pre-verbal, the penalty score is high. At the same time, when the model tends to predict post-verbal constituent with p=0.55, that would mean that model still thinks that post-verbal is a more likely option but now it is not as certain about it as in the previous example. Thus, it should be penalized less.

In case of the model presented in the main text, I only used test set (30% of the data) to quantify model's performance with log-loss function. It was quantified in two ways: (1) log-loss function by individual doculects (Figure 12); (2) log-loss function per entire dataset (Figure 13). Both Figures

Figure 16: Log-loss score function when position is constant and p varies.

(a) Always set at 0 (red)

(b) Always equal to 1 (blue).



also contain baseline scores. Computing baseline log-loss means that $\hat{p}$ in the equation 7 is replaced by a proportion of more frequent positions per doculect/dataset by the total number of data points per doculect/dataset. For instance, in case of Turkish of Ankara, we compute that proportion of pre-verbal to total number of constituents is around 93%. Hence, the baseline for Turkish Ankara is to fix value of $\hat{p}$ at 0.07 and compute average log-loss against the positions coded in WOWA. Given the way log loss baseline is computed, one can infer that more homogeneous doculects will have a lower baseline than more heterogeneous ones. The same is done for the entire dataset for the Figure 13. As the entire dataset is quite variable in terms of constituents' positions, baseline is set very high and is not that difficult to fit a model that would be better than that. Because of it, I find it more useful to use log-loss function applied per individual doculect.

WORD COUNT

15893