



Chapter 1

Post-predicate elements in the Western Asian Transition Zone: Data, theory, and methods

 Geoffrey Haig^a, Mohammad Rasekh-Mahand^b, Donald Stilo^c,  Laurentia Schreiber^a & Nils Schiborr^a

^aUniversity of Bamberg ^bBu-Ali Sina University Hamedan ^cMax Planck Institute, Leipzig (retired)

This chapter spells out the conceptual and methodological foundations for the volume, summarizes previous research, illustrates the methodology and analysis, and presents the results of two case studies. We provide evidence in support of a semantically finer-grained approach to word order that distinguishes between various non-subject constituents, and illustrate how this can be leveraged to detect areal effects in syntax. We implement this approach on a sample of language corpora from 35 languages, including Turkic, Iranian, Semitic, Hellenic, Kartvelian, and Armenian from what we term the Western Asian Transition Zone (WATZ). In a first case study, we demonstrate the existence of a robust Goals Last effect across the entire database and formulate a revised hierarchy for postverbal placement. Our approach identifies the specific properties of spatial goals that distinguish them from metaphorically related roles such as recipient, addressee, and benefactive, which previous studies had conflated. In a second case study, we investigate weight effects on post-verbal placement, concluding that overall, the impact of weight is minimal, a finding reflected in several chapters of the volume. The final section summarizes the contributions to the volume, and the Appendices provide raw data summaries across the entire WOVA data set, and information on sources.



1 Theoretical preliminaries

1.1 General background

This volume represents the collaborative outcome of a team of researchers, all of whom contributed expertise and data on languages of what we loosely refer to as the Western Asian Transition Zone (WATZ, cf. Section 1.2). The companion enterprise to this volume is a portfolio of online accessible, multiply-reusable digital resources, comprising of two data-sets: WOWA (*Word Order in Western Asia*), a multi-lingual corpus containing approximately 40 data sets from a sample of languages across WATZ (Haig et al. 2022), and HamBam (*The Hamedan-Bamberg Corpus of Contemporary spoken Persian*, Haig & Rasekh-Mahand 2022). HamBam is a richly annotated corpus of a single language, colloquial spoken Persian, based on the Multi-CAST architecture (Haig 2015a, Schnell et al. 2023).¹ It is designed for finer-grained investigations of word order, prosody, and register in a single language (Persian), while WOWA is a multi-lingual database designed to investigate the transition-zone phenomena outlined in the following paragraphs. Most of the research reported here is based on WOWA.

Our research is intended to satisfy the requirements of “reproducible research,” in the spirit of Berez-Kroeker et al. (2018), and the emphasis on accessibility and accountability of primary data, and maximal transparency of analysis procedures have been guiding principles throughout: research should be conducted in a manner “which allows readers to confirm claims about language structure through direct access to the original observational data” (Berez-Kroeker et al. 2018: 6). In this overview chapter, we introduce and exemplify the data sources, theoretical concepts and research questions, and illustrate the main findings with two case-studies. Figure 1 shows the location of the doculects in WOWA at the time of writing; an overview of all data sources is available in the Appendix to this chapter.

1.2 The Western Asian Transition Zone (WATZ)

The concept of “transition zone” has been discussed in various guises (e.g. “buffer zone” Stilo 2005, “intersection zone” Stilo 2009, or “typological sandwich” Szeto & Yurayong 2021, see Haig et al. In press). Here, we continue the terminology introduced in Haig & Khan (2019); we define a transition zone as a geographic area lying at the intersection of two contiguous regions characterized by diametrically opposing values for some linguistic feature. The choice of feature is essentially

¹For WOWA see <https://multicast.aspra.uni-bamberg.de/resources/wowa/>; for HamBam see <https://multicast.aspra.uni-bamberg.de/resources/hambam/>.

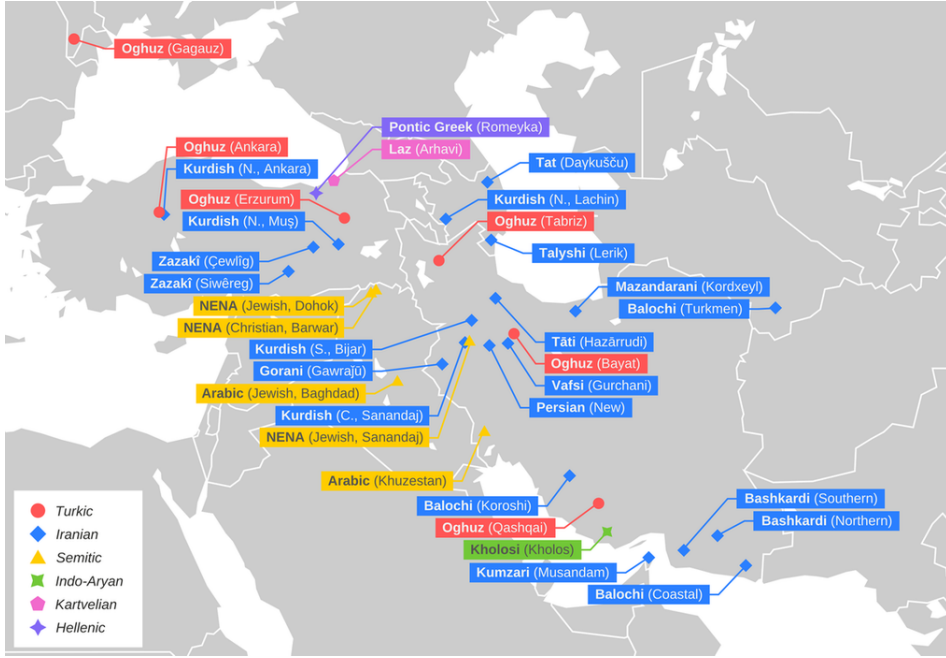


Figure 1: Locations of doctules in WOWA (November 2023)

unrestrained, and determined primarily by the research questions of the investigators, though considerations of general theoretical pertinence, availability of appropriate data, and operationalizability of the feature values concerned, will also play a role. It is clear that a transition zone, as just defined, can be identified at very different degrees of granularity: even a dialect isogloss, separating related dialects characterized by, for example, the presence versus absence of nasalized vowels, could be considered a transition zone, albeit at the micro-level of structural differentiation.

The Western Asian Transition Zone is at the other extreme of granularity. It is defined by the overlap of two areas of continental scale, which differ with regard to the following features (and a number of ancillary features, which are taken up at various points below): OV versus VO word order, and adpositional type (prepositions vs. postpositions, with some additional complications). An approximation of the global distribution of these two features can be found by considering the two maps in WALS (Dryer 2013b,c), Feature 83A (OV/VO) and Feature 85A (adposition order). On both maps, two adjacent macro-areas can be identified within Eurasia and the Indian Sub-Continent, each dominated by languages with opposing feature values: The first is the central and eastern Asian land block, dom-

inated by Robbeets' (2017) "Transeurasian languages" (Turkic, Mongolic, Tungusic, Japonic, Koreanic), which are uniformly characterized by OV and postpositions. These features flow seamlessly into the Indian Sub-Continent, dominated by Eastern Iranian, Dravidian, and Indo-Aryan. The second macro-area is to the southwest, where we find the VO and prepositional languages of North Africa, Western Europe and the Circum-Mediterranean region (Afro-Asiatic, Romance, Hellenic). The area of overlap between these two macro-areas is what we refer to as the Western Asian Transition Zone (WATZ). The core of WATZ is the lower catchment areas of the Euphrates and Tigris valleys, springing from the elevated plateau of today's Eastern Turkey and descending to the alluvial plains of northern Iraq and Syria. Today, this region is divided between four nation states, Turkey, Syria, Iraq and Iran.

Like most of the units that have been suggested in areal linguistics,² the geographic boundaries of WATZ cannot be precisely defined. Partly, this is due to the nature of a transition zone, which often implies fade-out phenomena (Stilo 2012), rather than abrupt transitions from one feature value to another. An example of such a fade-out phenomenon is the feature "order of spatial Goals relative to their governing predicate," as in a clause such as *the girl went to the market*. We can distinguish two options: the Goal precedes the verb (GV), or it follows the verb (VG). Mapping the frequencies of these two orders across the languages of WATZ reveals an increase in the frequency of Verb-Goal (VG) ordering as one progresses westwards. For example, Iranian languages with long-standing Semitic influence, from the southern and westernmost peripheries of WATZ, show almost 100% post-verbal Goals (in the sense of "Goals of movement"; see below on terminology), matching the figure that characterizes most of Semitic, and of the westernmost branches of Indo-European. Similarly, Turkic languages such as Qashqai, which are under heavy influence from western Iranian languages, also show high rates (>60%) of post-verbal Goals (Haig et al. In press). Note that both Iranian and Turkic are generally considered to be OV (sometimes erroneously equated with "verb final"), so the high frequency of post-verbal Goals is not expected for these languages. For OV languages situated further northward and eastward, the figures drop, for example in the Balochi variety of Turkmenistan (Nourzaei & Haig 2024 [this volume]), or Mazandarani of the Caspian region (Stilo & Haig 2022). Furthermore, initial counts from related OV languages further eastward beyond WATZ suggest they drop still further; in Indo-Iranian (Dardic) Kalasha

²Friedman & Joseph (2017: 75) discuss the issue of defining the boundaries of Sprachbünde, noting "the boundaries are as elastic as the micro-zones of convergence that add up to the larger convergence area." The boundaries, and even the set of languages and varieties involved, remain disputed even for intensely-researched linguistic areas such as the Balkan Sprachbund.

(Northern Pakistan, bordering Afghanistan), provisional counts of the texts in Petersen (2015) suggest rates of post-verbal Goals below 30%, while in the texts from Dukhan, a Turkic variety from Northern Mongolia, the figure approaches zero. If the basic hypothesis behind WATZ is correct, then we might predict similar low values of VG for unrelated OV languages toward the eastern fringe of Asia, such as spoken Japanese and Korean, but this remains to be tested.

Although our current data coverage is low-density and uneven, we tentatively hypothesize that in the OV languages of Asia, increasing rates of post-verbal Goals roughly correlate with increasing proximity to the Mesopotamian core of WATZ.³ But we can identify no precise geographical isogloss that constitutes a categorical border separating VG from GV. Rather, we are dealing with a continuum of values, which we assume extends beyond the region defined by the sample in Figure 1.

Returning to the broader theoretical interest of transition zones, it has been suggested that regions of intense contact (a hallmark of transition zones) are havens for typologically rare constructions. Harris & Campbell (1995: 137) note that certain word-order constellations only emerge in contact situations, and our data lend some credence to this view. Furthermore, transition zones are overall smaller than the macro-areas that engender them, and are therefore likely to contain fewer languages. The probability that any random language sample includes languages from transition zones is thus lower than the probability of selecting languages from within established linguistic areas. To this extent, the contributions to this volume are thus intended to counterbalance an existing bias in language sampling. To conclude, transition zones are not definable in terms of a precisely circumscribed geographic region. Rather, they should be seen as hypotheses which demarcate a potentially fruitful set of languages located at a region of conflicting feature values, which can serve as an experimental setting for investigating the broader question: what happens when languages with opposing feature values collide?

1.3 Word order

The term “word order” is often taken as synonymous for the traditional Greenbergian six-way typology (S/V/O). In our work, however, we follow Dryer (1997,

³It is worth noting that areality alone does not fully account for the findings; the phylogenies of the languages concerned are also relevant, with the Iranian languages apparently the most prone to areally-induced word-order variation; see Haig et al. (In press), Bickel (2017: 42) on the interplay of areality, inheritance, universal principles in co-determining language structure, and Haig & Schiborr (In review), for arguments in favour of a universal Goals Last principle.

2013a), who proposes decomposing the traditional six-way typology into two binary sub-features, S/V and V/O. Here we focus almost exclusively on the relative ordering of direct object and verb (V/O), while setting aside the position of subjects (S); see below for empirical justification, and Dryer (1997, 2013a) for the arguments against the six-way typology. In line with recent work in corpus-based typology (see Section 1.4), we extend the typology to include the position of other verbal arguments, such as Recipients, Locatives, copula complements, Goals, Addressees, relative to the verb. Our motivation for this is entirely data-driven: a large body of research (see Section 2) on the languages of Western Asia suggests that beyond direct objects, other less prominent and often overlooked constituent types provide sensitive indicators for contact influence (Haig et al. *In press*). Furthermore, the default assumption that the position of direct objects (i.e. OV vs. VO) can be generalized across other verbal arguments, is falsified in many languages of the region, which exhibit consistent OV order, but simultaneously have post-verbal placement of certain non-direct objects; see Section 4 below for an overview of the relevant findings from WOVA. Word order typology has tended to either ignore non-direct objects, or to subsume them under umbrella terms (e.g. ‘obliques,’ Hawkins 2008, Levshina 2019, Jing et al. 2021; or ‘PP’ in Frommer 1981). Our research indicates that a finer-grained semantic approach to non-direct-objects is more appropriate, which we spell out in Section 4 below.

A perennial debate in word order typology concerns how, or indeed whether, one can identify some kind of “basic word order” for a language. There are a number of issues at stake, which we will address here. First, although scholars such as Mithun (1992) have argued against the universality of basic word order, it is important to note that most researchers engaged in word order typology (see e.g. Dryer 2007 and Song 2018 for summaries) have never claimed that every language has a “basic word order.” Thus, the global overview of the VO vs. OV word-order parameter in (Dryer 2013c) assigns all languages in the sample to one of three types: OV, VO, and “no dominant order,” with the latter comprising some 7% of the 1,518 languages in Dryer’s (2013c) sample. It has always been acknowledged that it is not possible to identify a single order as “basic” for every single language. But that does not invalidate the enterprise of word order typology as a whole, any more than the fact that some languages do not have lexical tone invalidates a cross-linguistic approach to tone systems.

At least three different approaches to basic word order can be identified. First, the frequency of usage. Contrary to what is often claimed, a frequency approach does not necessarily imply a simple majority decision. Dryer (2007: 11) proposes that the basic (or dominant) order is that variant which is at least twice as frequent as the next most frequent order. In the case of a binary feature such as OV

vs. VO, that would mean one order would account for at least 66% of the relevant cases in order to count as the basic order. The frequency approach raises a host of methodological issues related to corpus size and representativity and ignores many finer nuances, some of which we take up below.

A second approach seeks to define a particular constructional sub-type which is taken as prototypical for the construction under consideration. For example, Siewierska (1988) provides a widely-cited rule-of-thumb for identifying the basic S/V/O order cross-linguistically. According to her, the basic order is that which:

[...] occurs in stylistically neutral, independent, indicative clauses with full noun phrase (NP) participants, where the subject is definite, agentive and human, the object is a definite semantic patient, and the verb represents an action, not a state or an event (Siewierska 1988: 8).

This would rule out, for example, interrogatives, subordinate clauses, or clauses in which either S or O is pronominal, and it would also exclude transitive clauses with verbs like ‘see’ or ‘know,’ which express states rather than actions and do not involve a ‘semantic patient.’ In principle, this is a reasonable approach and has gained some currency. However, we see no obvious justification for ruling out transitive clauses with indefinite direct objects, particularly as overall, direct objects are among the most likely argument types to host indefinite referents (Haig et al. 2021: 164, Schnell et al. 2023).

Another variant of the prototype approach is implemented by Bağrıaçık (2018) and Neocleous (2020), both investigating varieties of Asia Minor Greek.⁴ These authors focus on what is variously termed “pragmatically neutral” or “unmarked” transitive clauses, in which neither argument “is associated with either a topic, or a focus reading” (Bağrıaçık 2018: 150). Such clauses only occur under very specific discourse conditions, three of which are identified by Bağrıaçık as follows: (i) answer to an all focus question (*What happened?*); (ii) introductory clauses of narratives, where both subject and object are newly introduced into the world of discourse; (iii) generic statements (*the Earth orbits around the Sun, Sam knows Tibetan*), (Bağrıaçık 2018: 151–154). Neocleous (2020) adopts a similar approach, though with a different formalization of the concept of “pragmatically unmarked.” Note that Bağrıaçık (2018) concept of pragmatically neutral transitive clause is different from Siewierska’s (1988). Siewierska stipulates that in a

⁴Although both authors work within a Minimalist framework, thus approach “word order” from a rather different perspective to the typologically-oriented approach of Siewierska (1988), they nevertheless attempt to define a “basic” clause type on the basis of certain pragmatic and surface morphosyntactic properties, from which other orders are derived.

basic transitive clause, both subject and object are required to be definite, while Bağrıaçık's "introductory clauses," for example, imply that neither would be definite.

Along with prototype, and frequency-based approaches, there are two other connotations of "basic word order." The first is the concept of inherited, or historical word order. By this, we mean the word order that is reconstructible for the assumed proto-language of the languages under investigation. For example, there is little reason to doubt that both Turkic and Iranian languages had OV word order at the oldest period of attestation, or that proto-Semitic was VO, so these languages can be reasonably classified as historically OV and VO respectively. However, claiming that Semitic is historically VO does not equate to a claim that this is the "basic order" for all modern Semitic languages; basic order can change.⁵

Finally, a basic word order may be motivated by theory-internal considerations. This is the case for German, for which the basic word order is often claimed to be OV, with VO considered as secondarily derived via (some version of) verb movement. If we strictly applied the criteria of Bağrıaçık (2018), which relies on the concept of pragmatically neutral clause, or Neocleous (2020: 116), which invokes (among other things) word order in main declarative clauses, we would obtain a different result for German, because OV order in German is actually found in subordinate clauses. The fact that different criteria yield different results is reflected in the classification of German as "no dominant order" in Dryer (2013c).

Having briefly considered various interpretations of "basic word order," we turn to a methodological issue in connection with word order and small corpora. Both Siewierska (1988) and Bağrıaçık (2018) require a transitive clause to have two nominal (as opposed to pronominal) arguments. However, for research based on small corpora of spoken language, often without recourse to native speakers' judgements, this approach runs into an immediate problem. Cross-linguistically, in natural discourse, very few transitive clauses contain two overt lexical (as opposed to pronominal) arguments. Du Bois (2003: 62–63) provides data from five spoken language corpora indicating that the overall frequency of clauses with two lexical core arguments lies between two and seven percent, and similar findings are reported in the literature (see contributions in Du Bois 2003). Additional restrictions, such as requiring both S and O to be "pragmatically neutral"

⁵It may seem superfluous to labour this point, but it is nevertheless misunderstood in Asadpour (2022a: 42), who interprets references in the literature to "historical" and "inherited" word order in Neo-Aramaic as claims regarding "basic word order" in contemporary Neo-Aramaic. These are clearly separate claims.

(Bağrıaçık 2018), or both to be “definite” (Siewierska 1988), would further reduce the pool of valid tokens. It is no surprise that Bağrıaçık (2018) investigation of word order in Pharasiot Greek is largely informed by the elicitation of grammaticality judgements, rather than a quantitative analysis of naturalistic data.

Cross-linguistic research of spontaneous spoken discourse demonstrates that transitive subjects overwhelmingly express given information (>90%, Haig et al. 2021: 165), and are consequently predominantly either zero, or pronominal in form, rather than lexical NPs (cf. Du Bois’ 1987 ‘Avoid Lexical A’-constraint; see discussion in Haig & Schnell 2016). If we were obliged to exclude all clauses with pronominal or zero subjects, we would vastly decrease the number of potentially analysable tokens in the sample. For objects, however, the likelihood of lexical NP expression is very much higher, and the population of analysable tokens correspondingly larger. This is a further motivation to eschew the six-way SVO-typology, and to focus on the relative ordering of verb and direct object only. In the future, we may expand our investigation to include the position of subjects in the WOWA data, which are fully accessible and amenable for additional coding of subjects (see also Molin 2022, Rasekh-Mahand et al. 2024 [this volume], Forker 2024a [this volume],b [this volume]). Currently, however, we continue to work with the binary feature OV/VO.

1.4 Corpus-based approaches to word order

Methodologically, we apply a corpus-based typological approach (Wälchli 2009, Levshina 2019, Futrell et al. 2020, Gerdes et al. 2021, Haig et al. 2021, Schnell & Schiborr 2022, among many others). Within such an approach, the emphasis shifts away from assigning a ‘basic word order’ to a particular “language.” Rather, in corpus-based approaches, statements on word order refer to frequency distributions derived from actual corpora, and are probabilistic in nature. Strictly speaking, corpus-based approaches to word order yield a characterization of a specific corpus (a “doculect”), rather than “a language,” though we continue to use the over-simplified terminology here.

Corpus-based typological approaches to word-order typology are dominated by research on large, written corpora of languages with a pre-existing orthographic norm, for which copious quantities of pre-digitalized text are available (see in particular Universal Dependencies (UD)-consortium, Nivre et al. 2020).⁶ Consequently, there is a bias towards standardized (and mostly Eurasian) languages, and more importantly, towards written language. Cross-linguistic re-

⁶<https://universaldependencies.org/>

search based on spoken language corpora, on the other hand, is still in its infancy (see among others Schnell et al. 2021, Mettouchi & Vanhove 2021, Schnell & Schiborr 2022, Seifart et al. 2022, Levshina et al. 2023). Frommer's (1981) research had already demonstrated significant differences between formal written and spoken Persian, indicating that written language data cannot be assumed to reliably reflect the structures of spoken language. Any research agenda that purports to investigate the impacts of processing and production constraints on language structure would be well advised to focus on the mode of language production where these constraints are operative in real time – and that would not be written language (see Schnell et al. 2021 for summary arguments). In Section 5, we present an investigation into the role of weight as a predictor for word order, based on our spoken language corpora, which illustrates the importance of controlling for modality.

As mentioned above, we focus on various non-subject constituents, and their position relative to the governing predicate, for example direct object and verb, or Addressee and verb, and so on. Our database is thus designed to answer the following questions:

- (1) a. What is the probability that non-subject argument A, in doculect X, belonging to language family Y, spoken at location Z, occurs after its governing predicate?
- b. Which variables influence this probability?

From the answers to 1a we could infer word-order “types,” by setting some pre-defined quantitative boundaries. For example, Levshina (2019: 559) provisionally classifies a language (=corpus) with greater than 80% VO as “VO,” with less than 20% VO as “OV,” and 20–80% as “mixed.” However, this is a matter of heuristic interpretation of the raw data, rather than principled classification of “Type.” It should be obvious that two doculects with values of 79% and 81% VO respectively are not necessarily exemplars of fundamentally distinct types (see Wälchli 2009 on “data reduction typology”). The variables that were tested for question 1b are presented in Section 3.2. Questions 1a-1b; they can be investigated both at the level of individual doculects, or by applying appropriate statistical methods to the entire sample or some sub-set thereof. In Sections 4 and 5 below we present case studies for the impact of semantic role (Section 4), and of weight (Section 5). Having outlined the theoretical background and the research questions, we illustrate the structure of the data sets and the methodology in the following sections.

2 Previous research on Western Asia, and terminological issues

The assumptions and aims of the WOWA project were inspired by insights gained over many years of previous research, and it is appropriate to briefly outline the main currents of that research. Typologists have long been aware of the “mixed typology” of Iranian languages, e.g. Comrie (1989: 19) on Persian, a language with OV in the VP, but head-initial NPs, prepositions, and clause-initial complementizers (see Dabir Moghaddam 2018 for a recent summary). Don Stilo developed the idea that the mixed typology of Persian was shared to differing degrees by other West Iranian languages, and that the degree and nature of West Iranian mixed typologies followed an approximate areal distribution. Stilo’s claim was that West Iranian was sandwiched between the opposing typologies of Semitic (consistently head-initial) and Turkic (head-final), with different West Iranian languages synchronizing with the profile of their respective geographic neighbours. These ideas were fleshed out with a survey of adpositional types in Stilo (2005, 2006, 2009), and developed in a number of other publications (2012, 2018b, 2018a).

Frommer (1981) noted a further non-harmonic aspect of West Iranian syntax: the post-verbal positioning of certain kinds of non-direct-object arguments. Frommer (1981) focussed on the syntax of “informal Persian” (IP), including both spoken and written samples from different registers. This was, in fact, the crucial breakthrough: formal written Persian, the more usual object of study, is rather consistently “verb final,” hence post-verbal elements are a fringe phenomenon that had not been systematically investigated.⁷ Frommer (1981) was the first systematic analysis of post-predicate elements in different registers of informal Persian;⁸ his findings can be summed up as follows: (i) across the different registers of informal Persian, there is a cline of formality such that a lower degree of formality correlates with an increase in post-predicate elements; in-group domestic conversational Persian exhibited the highest levels; (ii) semantic role is crucial,

⁷Lazard (1957) had already noted the predominance of post-verbal Goals in informal spoken Persian, but did not systematically investigate the topic.

⁸Frommer did not actually investigate formal written Persian, and we still lack a systematic study. Parizadeh’s (Parizadeh & Rasekh-Mahand 2024 [this volume]) study of Early Classical New Persian (11–14th Century CE) demonstrates near 100% verb finality in these written texts, which matches the native speaker intuition of one of our authors regarding contemporary formal written (e.g. academic prose) Persian. More recent corpus-based approaches to written Persian (e.g. Faghiri et al. 2018) investigate the relative ordering of pre-verbal constituents, while post-verbal constituents lie outside the purview of this research. Formal written Persian is thus essentially considered to be a “verb final” language.

with goals of motion (“destinations” in Frommer’s terminology) as the leaders in post-verbal placement, across all registers; (iii) information status is relevant for post-verbal placement (focal versus non-focal) of direct objects, but appears to be irrelevant for goals of motion; (iv) there is a stronger tendency for post-verbal constituents to lack overt flagging. By and large, these findings have been confirmed on more recent corpora of spoken Persian (Rasekh-Mahand et al. 2024 [this volume]).

From a comparative Iranian perspective, the remarkable aspect of Frommer’s findings is that they closely align with findings from the lesser-researched and generally non-standardized West Iranian languages documented in this volume and elsewhere. What this suggests is that the phenomena which Frommer identified were not merely irregularities specific to informal spoken Persian, but in fact reflected word order traits of considerable antiquity, which characterize most (perhaps all) West Iranian languages (see Nourzaei & Haig 2024 [this volume], Korn 2024 [this volume], Nourzaei 2024 [this volume], Rasekh-Mahand et al. 2024 [this volume], Parizadeh & Rasekh-Mahand 2024 [this volume], Mohamadirad 2024 [this volume]). From this perspective, it is the strictly verb-final, formal written Persian that is the exception when it comes to West Iranian word order. This has considerable implications for the diachronic study of word order, which is largely reliant on written language sources.

Research on post-predicate elements in other Iranian languages began with Kurdish (Haig 2015b, Haig & Thiele 2014, Haig 2022d), and has since expanded to neighbouring languages (Haig 2015a, 2017, Stilo 2018a, Jahani 2018, Asadpour 2022a, and contributions to this volume). Most of this research is based on corpora of spoken narrative texts (see below), though increasingly enhanced with experimental data (see Skopeteas 2024 [this volume]). It has emerged that while all West Iranian languages investigated to date (with the exception of Kumzari, Anonby 2015) are consistently OV (see below), like spoken Persian, they are not “verb final” because a significant number of non-direct object arguments regularly follow the verb. A similar pattern can also be observed in Turkic languages in contact with Iranian (Schreiber et al. 2021, Stilo 2021a). A second point that quickly emerged from the earlier studies is that the nature, and systematicity, of post-verbal arguments follows an approximate areal distribution, along the lines of Stilo’s (2009) suggestions. Among Iranian OV languages, the highest frequencies, and greatest variety of post-verbal argument types, are attested among varieties of northern Kurdish spoken in Iraqi Kurdistan and adjacent regions of southeastern Turkey, Syria and Iran, a region we provisionally refer to as Mesopotamia. Mesopotamia is of course also home to a number of historically VO Semitic languages, (Neo-)Aramaic and Arabic, which have co-existed with

Kurdish and other OV languages for centuries, and indeed for millennia in the case of Aramaic. In these VO languages, it is universally the case that other non-direct object arguments also follow the verb, and it seems plausible to assume that the syntax of these languages had some impact on the Iranian languages with which they shared territory for at least 2000 years, and ultimately also on Turkic (perhaps via Iranian in many cases).

In a pilot study, Haig & Thiele (2014) compared word order in naturalistic texts from a sample of languages mostly from Mesopotamia.⁹ The authors identified four types of arguments that are predominantly post-verbal in these languages, cited in the original formulation as follows:

- Recipients of verbs of transferred possession (e.g. GIVE)
- Destination or direction of verbs of movement (e.g. GO, RUN, FALL)
- Destination or direction of verbs of caused motion
(e.g. PUT, PLACE, TAKE)
- Addressees of verbs of speech (e.g. SAY, SPEAK, PROMISE)

Examples illustrating these four types, from Badini Kurdish (from the Gulli and Akre dialects of Iraqi Kurdistan, from Haig & Thiele 2014, citing MacKenzie 1962) are provided in (2-5):

(2) Recipient

Northern Kurdish Akre (MacKenzie 1962)

min kič-ā xo dā ta

1SG.OBL daughter-EZ.F REFL give.PST.3SG 2SG.OBL

‘I have given my daughter **to you**.’

(3) Addressee

Northern Kurdish Akre (MacKenzie 1962)

sultān-ī got-a ahmad halwāčī

Sultan-OBL.M say.PST.3SG-DRCT Ahmad Halwachi

‘the Sultan said **to Ahmad Halwachi**.’

⁹The varieties were Northern Kurdish from Iraqi Kurdistan, Northern Kurdish from Midyat, Southeastern Turkey, Northern Kurdish from Muş and Erzurum; Northeastern Neo-Aramaic (Jewish) from Urmi, West Iran, and from Koy Sanjaq, Iraqi Kurdistan, and Turkish from Erzurum, Turkey. For comparison, they also included corpus data from a dominant VO language, Cypriot Greek. It was already apparent from this small data set that outside of Mesopotamia, Recipients, Addressees, and Goals of motion do not necessarily pattern alike.

(4) Goal of simple motion

Northern Kurdish Akre (MacKenzie 1962)

harduk rābon, hāt-in-a bāžar-ī

both get_up.PST.3PL come.PST-3PL-DRCT town-M.OBL

‘Both of them got up and came **to the town**.’

(5) Goal of caused motion

Northern Kurdish Gullī (MacKenzie 1962)

kir t=sēnīk-ā dayk-ā xwa dā

do.PST.3SG ADP=tray-F.EZ mother-F.EZ REFL ADP

‘(He) put (it) **on his mother’s tray**.’

In the Mesopotamian languages investigated in Haig & Thiele (2014), all four semantic types exhibited broadly similar rates of post-verbal placement, which motivated the authors to define a macro-role, labeled “Goal,” that would encompass all four types (and some further types such as final state of a change-of-state verb, see below). In retrospect, this terminological decision proved injudicious, for two reasons. First, it introduced ambiguity to the term “Goal,” which could either be understood in the narrower sense of “Goal of verb of motion,” or in the broader sense that would include Recipient, Addressee, etc. Second, it has become increasingly evident that many languages of WATZ do not lump Addressees, Recipients, and Goals of verbs of motion together (Section 4 below for data), thus casting doubt on the validity of a macro-category altogether. A broadly similar macro-category was subsequently adopted by Asadpour (2022a,b), who relabels it as “Target,” and this terminology has been used in the contributions to Asadpour & Jügel (2022). While the re-labeling alleviates the ambiguity problem, it does not resolve the empirical problem that outside of some varieties of Kurdish with deep historical ties to Semitic languages, Addressees, Recipients, and spatial Goals do not pattern alike among the OV languages of WATZ, so the motivation for assuming a priori a meta-category is questionable.¹⁰ In an effort to restore clarity, we therefore eschew the macro-category sense of “Goal” in this volume, reserving the term “Goal” strictly in the sense of “Goal or endpoint of a predicate of motion or caused motion.” See end of Section 4.1 for discussion of “Recipient” vs. “Goal,” and (11) for an overview of roles distinguished in WOWA.

One of the observations in the earlier literature concerned the syntax of ‘final state’ constituents, defined here as expressions indicating the final state of a

¹⁰There is also a lack of consensus about the nature and number of categories that are included under “Target”; some researchers include variously Benefactives, and Final States of change-of-state predicates, rendering comparison across different publications difficult.

change-of-state (‘become’) predicate. Haig (2017, 2022d) noted that in much of Kurdish, the final states are significantly more likely to be post-verbal than the complements of copular expressions that do not imply a change of state, even when the lexical verb is the same;¹¹ compare (6) (change-of-state) and (7) (static state).

- (6) Central Kurdish Sanandaj (Mohammadirad 2022b: I, 1016)

bū-m=a wirdafirūš

be.SUBJ-1SG=DRCT peddler

‘(I will) become (a) peddler.’

- (7) Southern Kurdish Bijar (Mohammadirad 2022c: D, 0282)

aware kur bī ...

if boy be.SUBJ.3SG

‘If it were a boy ...’

Similar phenomena have been noted for unrelated OV languages in close contact with Kurdish. For example, in the Northeastern Neo-Aramaic (NENA) dialect of the Jewish speech community from Urmi (West Iran), “the complement of the verb *qlb* ‘turn into’ is invariably placed after it” (Khan 2008a: 323). Although post-posing of complements of change-of-state predicates is widespread in the region, it is not grammaticalized to the same extent in all languages. Having outlined some of the main currents in earlier research and clarified terminology, in the remaining sections, we describe the design of the database and present two case studies illustrating cross-corpus results.

3 Design of the WOWA (Word Order in Western Asia) database

The WOWA sample includes data sets from 35 languages and varieties, based on monological, unscripted, spoken texts (see Figure 1 and the Appendices for details). As most of the project was conducted during the 2020–2022 pandemic, it was not possible to systematically select locations and languages in which to conduct dedicated fieldwork; rather, we have been obliged to rely on pre-existing

¹¹Interestingly, post-verbal placement of a change-of-state complement is much more likely when the complement is nominal (e.g. ‘she became (a) teacher’), rather than adjectival (e.g. ‘she became rich’).

resources. The result is that we were unable to compile a geographically or phylogenetically balanced sample of varieties. Nevertheless, the present sample represents the largest and most systematic data source currently available for investigating word order across the region.

The data sets stem from a range of distinct research contexts. Some are based on texts extracted from the published output of scholars working within individual philologies (e.g. the Neo-Aramaic texts of Barwar, Northern Iraq, originally published in Khan 2008b, a sub-set of which is analysed for a WOWA data set, Stilo 2021a). Other data-sets are taken from published sources of national language academies in the framework of dialect surveys (e.g. the Erzurum dialect of Turkish, which feed into Dogan 2021a), while others stem from contemporary language documentation projects, such as the Hazarrudi Tat texts used in Izadifar (2022) and the Qashqai texts in Schreiber (2021a).

For most data sets, the most widely represented genre is traditional narrative, but some data sets also include stimulus-based narratives (e.g. Pear story (Chafe 1980) retellings). The texts have been transcribed according to the academic tradition of the original researcher (we have not attempted to impose a common transcription scheme), and translated into English (in one case, into German). Generally, each data set includes more than one text, in most cases from different speakers; the composition of each data set is described in the accompanying metadata, and the source of each token (i.e. the individual text, and speaker) is recoverable. The main criteria for inclusion of a dataset in WOWA are a minimum yield of 500 codable tokens, reliable and authentic spoken data, and no restrictions on data accessibility.

3.1 Content of each data set

The list below provides the downloadable resource types included in WOWA. The first three are available for all data sets, while the other three are accessible to varying degrees, depending on the nature of the source data:

1. All files: Complete data set in a single ZIP-directory.
2. Coded values: The actual coded data (see below), in Excel and TSV format.
3. Metadata: A text document containing information on sources, references, speaker metadata, links, and other relevant information.

4. Source texts: Contains an orthographic rendering of the entire text with a translation, often from a published source, or provided by the contributor. In some cases, the source texts include additional morphological glossing or other information.
5. Sound files: The original sound files (where available), in .WAV and .MP3 formats.

3.2 Segmentation and token coding

The basic units of the database are tokens of prosodically independent (rather than bound), referential, non-subject constituents. Creating the database thus involves identifying the relevant tokens, and coding them for a series of features (see below). Texts selected for inclusion into the corpus are first segmented into strings that correspond approximately to meaningful utterances (in many cases this corresponds to a clause), termed utterance units. Each utterance unit is accompanied by a translation into English (column “utterance_translation” in Table 1 below). Utterance units are consecutively numbered and entered as single rows in the database, initially implemented in an Excel spreadsheet.

In a second step, all tokens of referential, non-subject constituents are identified and entered into a distinct cell (token) aligned with its source utterance unit. Note that clausal constituents (complement clauses, etc.) are not included as tokens. If an utterance unit contains more than one relevant token, that row of the data is repeated. If an utterance unit contains no relevant token (for example, a simple intransitive clause often does not contain any overt non-subject constituent, see 0006 in Table 1 below), then the token column remains empty. Note that these non-coded utterances remain in the data set, which thus preserves the overall unity of the original text, and maintains its re-usability for future research.

The basic structure is illustrated in Table 1, from the Hazarrudi Tat data set (Izadifar 2022). The utterance in 0006 does not contain a relevant token, thus the token columns are empty. The utterance in 0007, on the other hand, contains two relevant tokens (‘some prey,’ and ‘there’). Each receives its own ID (0007 and 0008), and the utterance unit is repeated, enabling tokens to be systematically associated with their contexts across all analysis steps. WOVA currently contains approximately 20,000 analyzed tokens in context.

Table 1: Fragment of Hazarrudi Tat data set (Izadifar 2022)

token ID	utterance unit	utterance translation	token	token_translation
0001	<i>bale čemā rustā de i nefar ve</i>	yes, there was a person in our village	<i>čemā rustā de</i>	in our village
0002	<i>šekārči ve</i>	he was a hunter	<i>šekārči</i>	hunter
0003	<i>ševi šekār</i>	he had gone for hunting	<i>šekār</i>	hunting
0004	<i>ševi šekār čemā kua de</i>	he had gone hunting in our mountains	<i>šekār</i>	hunting
0005	<i>ševi šekār čemā kua de</i>	he had gone hunting in our mountains	<i>čemā kua de</i>	in our mountains
0006	<i>i jangali ve</i>	there was a forest	(no token)	
0007	<i>berā de i šekāri bezzeše</i>	he killed (hit) some prey there	<i>i šekāri</i>	some prey
0008	<i>berā de i šekāri bezzeše</i>	he killed (hit) some prey there	<i>berā de</i>	there

Once identified, each token is coded for a number of features, which fall into the following three types:

Doculect-related features:

- Genetic affiliation (e.g. Iranian, southwestern)
- Doculect location (latitude, longitude)

Context-related features

- Text and speaker identification (unique identifiers are assigned, which are described in the accompanying metadata document)

Linguistic features of the token and immediate context

- Classifiable versus non-classifiable (if an utterance unit contains either no relevant token or none that can be unambiguously classified). Non-classifiable tokens are not included in statistical analyses
- Pronominal versus nominal form
- Animacy
- Definiteness (only applied to direct objects)
- Weight
- Role (see (11))
- Flag (adposition, or case-marking)
- Position relative to the governing predicate (the dependent variable): before (0) vs. after (1)
- Comments (free text entry)

Obviously, the set of linguistic features could easily be extended to include, for example, main versus subordinate clause, finer-grained metrics of topicality, and so on. The final decision on which features to include was a compromise determined by the partially conflicting demands of theoretical relevance, and practical concerns such as economy of time and resources, simplicity of implementation across multiple languages with multiple coders, and replicability and transparency of coding-decisions. Previous research has pointed to the importance of pronominal versus nominal (e.g. Gerdes et al. 2021), animacy, weight (references in Section 5), informativity (Faghiri & Samvelian 2020), flagging and role (see Section 2 above), and these are also features that best satisfy the practical constraints just mentioned. Note that the raw data are available for coding additional features in the future. For each linguistic feature, coders select from a pre-defined set of options, which are explained in the Coding Guidelines.¹²

Coders work with the project coordinators, and problematic issues are resolved collaboratively to maximize cross-coder consistency. The coding scheme was presented and discussed collectively at two workshops (2019, 2020), and continued to evolve over the course of the project, before a final version was adopted in 2020. It should be evident that in a project of this nature, with multiple contributors working on multiple languages, compromise is inevitable. We have strived

¹²https://multicast.aspra.uni-bamberg.de/resources/wowa/data/_docs/guidelines/wowa_coding-guidelines.pdf

to maintain the fine line between maximal simplicity and generality (limiting the number of coding options), while maintaining sufficient flexibility for capturing the range of cross-language variation contained in the data. Nevertheless, some degree of coding indeterminacy is inevitable, and for this reason, we include the coding option “other” in all linguistic categories to capture those instances where the analyst cannot decide among the available options. The full list of coding options is available in the Coding Guidelines; by way of illustration, we demonstrate in Table 2 the linguistic coding of the eight items from Table 1 above.

Table 2: Coding the linguistic values for the tokens in Table 1

token	token translation	pro	anim	weight	weight2	role	flag	position
<i>čemā rustā de</i>	in our village		inan	2	11	loc	postp	0
<i>šekārči</i>	hunter		hum	1	7	cop	bare	0
<i>šekār</i>	hunting		inan	1	5	Goal	bare	1
<i>šekār</i>	hunting		inan	1	5	Goal	bare	1
<i>čemā kua de</i>	in our mountains		inan	2	9	loc	postp	1
(no token)								
<i>i šekāri</i>	a prey		inan	2	7	do	bare	0
<i>berā de</i>	there		adv	1	6	loc	postp	0

The “pro” column is empty in Table 2, because there are no pronominal tokens in this stretch of discourse. The “weight” column records orthographic words, except function words solely employed as flagging devices (e.g. simple adpositions). The “weight2” column is a finer-grained weight metric that is automatically generated, based on the number of characters contained in the transcription of the token (thus the first token consists of 11 characters); it provides a rough proxy for the number of phonological segments in each token. The column “position” is the dependent variable, and offers a binary option of <0> (pre-verbal) versus <1> (post-verbal).

The proprietary spreadsheet format used for data entry was dictated by practical considerations; most contributors use MS Excel (or equivalent) and were able to enter their data into the template that we provided. For the actual analysis, data are exported to R, a powerful and flexible programming language and platform for statistical computing.

With regard to the pronoun category, we have included only prosodically independent pronouns as tokens. For languages that make extensive use of clitic object pronouns, this means that the number of classifiable object tokens in these languages may be very low, which has a detrimental impact on the statistical analysis (authors have the option of noting the presence of clitic pronouns in the comments column (e.g. Schreiber 2021b), so the information is available for future analyses.) There are sound empirical reasons for distinguishing free and clitic pronouns, illustrated below from spoken Persian (Rasekh-Mahand et al. 2024 [this volume]): around 95% of nominal direct objects precede the verb (OV), as in (8). Clitic object pronouns, on the other hand, frequently right-attach to the verb, and indeed must do so if the verb is the sole available host, as in (9) and (10):

- (8) Colloquial New Persian (Izadi 2022: C, 0263)

doz=râ bord bâlâ

dosage=ACC carry.PST.3SG upwards

‘(He) increased **the dosage**.’

- (9) Colloquial New Persian (Izadi 2022: V, 2375)

mi-šenâs-im=ešân

INDIC-know.PRS-1PL=3PL

‘We know **them**.’

- (10) Colloquial New Persian (Haig & Rasekh-Mahand 2022:

oh_f_accident_0166)

be-bar-id=aš

IMPER-take.PRS-2PL=3SG

‘Take **him**!’

It would make little sense to count constructions such as (9) and (10) as ‘VO,’ apparently in contrast to the OV of (8). Examples (9) and (10) illustrate a language-specific rule of cliticization, which permits no variability of object placement in these examples. Clitic placement is a fascinating issue in its own right but of limited relevance for the principles operating in the linearization of independent phrases in syntax. Consequently, clitic pronouns hosted by the predicate are not included in calculations of pre- versus post-verbal argument placement. Clitic pronouns hosted by an item distinct from the predicate, on the other hand, are coded as “bound,” and the normal coding procedures applied. Depending on the analysis, pronominal tokens may be filtered out of a given sample.

4 The impact of semantic role: the “Goals Last” effect

4.1 Background

For the majority of languages in the sample, the variable “Role” turned out to be the most influential factor in determining pre- versus post-verbal position. The category “Role” in WOVA distinguishes the 19 categories shown in (11).

- (11) Role categories recognized in WOVA (see Coding Guidelines, Section 3.2)
- ABL** source of motion (‘she came out of the house’)
 - ADDR** addressee of a verb of speech (‘they spoke to him/asked her/begged the King’)
 - BECM** ‘become,’ i.e. the final state of a change-of-state (inchoative), predicate, such as ‘become X,’ ‘turn into X’
 - BECM-C** final state of a caused change-of-state predicate (‘they made him King,’ ‘she turned him to stone’)
 - BEN** benefactive; a person who benefits, or is disadvantaged, by an event without being directly impinged on by the action
 - COM** comitative; a person who accompanies another participant in some action, or state (‘I went to the market with my father’)
 - COP** complement of a copular expression (‘they were farmers’)
 - COP-LOC** locational complement of a copular expression (‘she was in the car’)
 - DO** direct object, which needs to be identified on language-specific criteria such as typical case marking properties
 - DO-DEF** definite direct object (which will include most pronouns), i.e. an item whose identity is recoverable from the context through previous mention or assumed deictic reference (‘she took that cup’)
 - GOAL** endpoint or destination of a verb of motion (‘it fell on the table’)
 - GOAL-C** endpoint or destination of a verb of caused motion (‘he put it on the table’)
 - INSTR** instrument for carrying out an action
 - LOC** static location (with no implication of movement) of a participant or event
 - OTHER** none of the available categories
 - POSS** possessed in a clause expressing possession ‘she had two brothers,’ unless the language has a HAVE verb and expresses the possessed in the same way as a direct object (do)

- REC** recipient of a theme in an event of transfer, typically GIVE
- REC-BEN** recipient-benefactive. This is included for contexts in which it is unclear whether a particular token is the recipient, or a benefactive of an action ('he bought the apples for us' – recipient or benefactive?)
- STIM** stimulus, typically of verbs of emotion, perception, desire – if they are not coded as direct objects (English 'she was afraid of the snake' (stim), but not 'she hates snakes' (coded as <do>))

The data reveal very divergent token frequencies of different roles. In fact, some are so infrequent that they offer little leverage for statistical purposes. Figure 2 below provides the respective proportions of different roles, whereby we have lumped together those role types that occur only marginally.

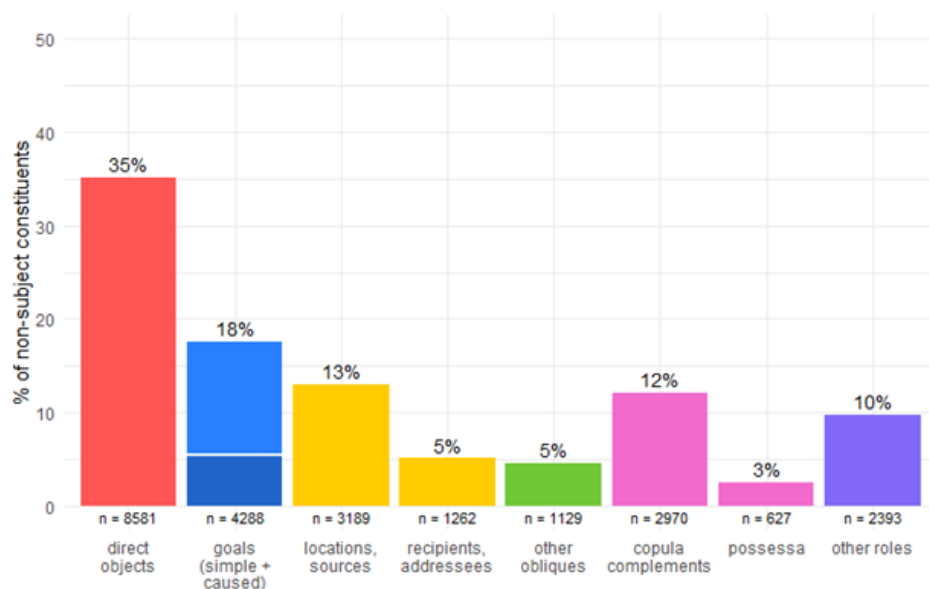


Figure 2: The respective frequency of non-subject roles across the WOVA sample

The most frequent role type in our data are direct objects, which account for around one third of all tokens. The next most frequent are goals of motion, including both simple motion and caused motion (distinguished in Figure 2 by different shades of blue). Recipients and addressees are relatively infrequent, and are therefore lumped together here (see below). Similarly, roles such as instrument, comitative, stimulus, and benefactive do not occur in sufficient numbers

to permit meaningful quantitative analysis, so have been combined under “other obliques.” The “other role” category includes tokens that were not classifiable under any of the available role-categories.

Turning now to the respective frequencies of post-verbal placement, Table 3 visualizes the general trend discernible across the data set. Four categories are distinguished: Goals (including caused goals), Recipient+Addressee, other obliques, and direct objects. For each of those four categories, we have colour-coded any frequency values for post-verbal placement which exceed 66%.

The data in Table 3 permit a number of generalizations, which to our knowledge have hitherto not been recognized. First of all, the data underscore the exceptional status of spatial Goals when compared to any other role. More than two thirds of the sample languages have dominant post-verbal placement of Goals (>66%), and for the majority of these languages, dominant post-verbal placement is restricted only to Goals. In several of these languages, the frequency of post-verbal Goals is twice as high as the frequency of post-posing any other constituent. In other words, Goals are different. We can formulate this as a potential universal in (12):

- (12) If a language postpones any role with a greater than two-thirds frequency, it will postpone Goals.

Note that Table 3 combines the two roles Recipient and Addressee, due to the low absolute numbers of tokens in these categories. However, this obscures the fact that Recipients and Addressees do not actually pattern consistently across all data sets. In fact, three distinct patterns can be identified in the sample, bearing in mind that the relevant absolute numbers are small:

- (i) both Addressee and Recipient occur before the verb (e.g. Oghuz Erzurum, Zazaki Cewlig, modern spoken Persian);
- (ii) both occur after the verb (e.g. NENA Jewish Duhok, Central Kurdish Sanandaj, Kumzari);
- (iii) Recipients occur after the verb, but Addressees before the verb (e.g. Vafsi Gurchani, Northern Kurdish Muš, or Armenian of Agulis).

Crucially, no language is attested in the sample that has post-verbal Addressees, but pre-verbal Recipients. Bearing this in mind, we can formulate the following

1 Post-predicate elements in the Western Asian Transition Zone

Table 3: Frequency of post-verbal placement for different roles; shaded values indicate frequencies above 66%. Goals and direct objects include only nominal expressions; Recipients, Addressees, and other oblique roles (Locations, Sources, Instrumentals, Benefactives, and Comitatives) also include pronominal expressions. Two doculects with less than 8 observations in each role category are excluded. See Appendix B for raw data. (NENA = Northeastern Neo-Aramaic).

Doculect	% post-verbal			
	Goals	Recipients + Addressees	Other Obliques	Direct Objects
✿ Laz (Arhavi)	4	6	1	3
◆ Persian (New, Early Classical)	5	0	9	2
● Oghuz (Ankara)	7	19	7	2
● Oghuz (Erzurum)	38	8	9	7
◆ Balochi (Turkmen)	48	14	6	2
◆ Kurdish (Northern, Ankara)	59	13	6	0
✱ Kholosi (Kholos)	62	39	1	1
◆ Mazandarani (Kordxeyl)	63	12	8	3
◆ Balochi (Coastal)	63	28	18	7
◆ Bashkardi (Northern)	63	56	31	27
● Oghuz (Bayat)	64	43	13	4
◆ Zazaki (Cewlig)	91	15	0	5
● Oghuz (Tabriz)	75	21	13	1
◆ Persian (New)	84	25	19	5
◆ Kurdish (Southern, Bijar)	97	27	11	2
◆ Tat (Daykusecu)	78	34	30	15
◆ Kurdish (Northern, Muş)	89	36	5	3
◆ Bashkardi (Southern)	80	38	36	11
◆ Vafsi (Gurchani)	88	38	11	2
◆ Kurdish (Northern, Lachin)	81	40	4	2
◆ Balochi (Koroshi)	90	42	27	2
◆ Talyshi (Lerik)	72	45	26	2
● Oghuz (Gagauz)	73	47	38	51
◆ Zazaki (Siwereg)	100	52	8	5
◆ Tati (Hazarrudi)	92	53	16	3
● Oghuz (Qashqai)	71	88	16	8
◆ Kurdish (Central, Sanandaj)	94	95	22	1
◆ Gorani (Gawraju)	96	82	36	5
▲ NENA (Jewish, Sanandaj)	92	81	40	5
◆ Kumzari (Musandam)	100	97	87	7
▲ NENA (Christian, Barwar)	96	100	74	83
▲ NENA (Jewish, Dohok)	99	98	93	90
▲ Arabic (Jewish, Baghdad)	100	100	85	97

implicational hierarchy indicating the conditions under which a particular role type may occur as dominant (at least 66%) postverbal.¹³

(13) Goals > Recipients > Addressees > Other > Direct object

The hierarchy in (13) is to be understood as an implicational universal, which can be formulated as follows: based on token frequency in corpora of spontaneous spoken language, and frequencies of nominal as opposed to pronominal constituents, if a language postposes any of the roles in (12) with greater than two-thirds frequency (dominant), then it will also dominantly postpose all higher roles on the hierarchy. Thus for the languages in the WATZ sample – regardless of genetic affiliation – there is no language that has, for example, dominant post-verbal Addressees, but not dominant post-verbal Goals. There are, however, a considerable number that only have dominant post-verbal Goals, and no other roles. Finally, if any language has dominant post-verbal direct objects, then all other roles are likewise dominant post-verbal. Thus there are no languages that, for example, combine dominant post-verbal objects with pre-verbal Goals.

Currently it is impossible to say with certainty whether the regularities illustrated in Table 3 and expressed in (12) and (13) represent a peculiarity of the languages in and around WATZ, or whether they reflect a deeper trait of connected spoken language, which should surface in spoken-language corpora of any language. We suggest there are grounds for assuming that (13) does reflect, at least in part, universal tendencies. Haig & Schiborr (In review) compare the two ends of the hierarchy (direct objects and Goals) across a more extensive sample of spoken language corpora and report that in no corpus do frequencies of postverbal objects exceed frequencies of postverbal Goals, regardless of language type or genetic affiliation.

Even if only parts of (13) should turn out to be valid outside of WATZ, this would have considerable implications for understanding, for example, diachronic change in word order. Essentially, (13) predicts an ordered sequence in a shift from OV to VO and vice versa. If a language is, for example, consistently verb-final (i.e. no role is dominant post-verbal), then (13) predicts that any change towards less verb-finality would occur first with Goals, and proceed down the

¹³The hierarchy in (13) is similar to others that have been formulated in the literature (e.g. Frommer 1981, Haig & Khan 2019, Stilo 2018b), some of which add additional roles, or employ somewhat different terminology. One role that has often been included is Benefactive; however, our data for this role are sparse, and were unfortunately not always consistently coded, rendering interpretation of the results difficult. Currently they are included under ‘other’; this requires more research.

hierarchy, with direct objects the last to shift. For the other direction, i.e. a language that is consistently verb-initial in the VP (like English), the prediction is that if any argument type shifts across the verb, it will be direct objects first, and Goals last. There is thus an asymmetry in the way VO and OV languages can be expected to move closer to one another. Preliminary observation of word-order change in WATZ suggests that this holds, regardless of whether the shift is considered internally motivated, or contact induced.

At this point, the relationship between recipient and Goal roles merits discussion. In earlier work (Haig 2022d), it was argued, on the basis of Kurdish data, that these roles share a common semantic component, defined as “event endpoint,” which motivated the shared post-verbal syntax in Kurdish. However, as we have seen, for the majority of other languages in our sample, recipients and goals do not pattern alike. On the assumption that both share endpoint semantics, the question arises as to what inhibits post-verbal placement of recipients? To understand this, it is important to recall that word order is the product of competing motivations, of which iconicity is but one. These include verb-object adjacency (the tendency for direct objects and verbs not to be separated by other constituents), weight, animacy, and agency considerations. Thus word order in any given context is the product of multiple factors, including information structure, semantics, and configurational constraints. Recipients differ strikingly from goals in several dimensions relevant here: they are overwhelmingly human, with high frequency pronominal and first or second person, and are treated syntactically as direct objects in many languages (Haspelmath 2015). Thus we suggest that endpoint semantics are simply overridden by other factors in the ordering of recipients. The distinction between goals of verbs of motion, and recipients is clearly maintained in Haspelmath’s (2015) concept of “ditransitive construction,” which presupposes an element of transfer of possession, while goals of verbs of caused-motion, such as ‘put,’ lack such an entailment and are thus outside the purview of the typology of ditransitive constructions. We might add that conversely, transfer of possession does not necessarily entail movement: it is possible to give someone a house, or a piece of land, which involves no actual change of location of the “theme.” Our revised conclusion is thus that although shared endpoint semantics mean goals and recipients may pattern alike in some languages, the overall weight of evidence suggests that a distinction should be maintained (see Haig & Schiborr *In review* for a more detailed discussion).

5 The impact of weight on post-posing

5.1 Background

It is fair to say that in both experimental and corpus-based approaches to word-order typology, considerations of weight (however formalized) have attracted more attention than any other single factor (e.g. Faghiri & Samvelian 2020, Schnell & Schiborr 2022, Wasow 2022: 5–10). However, as Yao (2018) points out, most of the relevant research considers weight as a factor in determining the relative order of **constituents occurring on the same side of the predicate**, for example the relative ordering of the two PP’s in (14) (from Wasow 2022: 6).

- (14) *The gamekeeper looked [through his binoculars] [into the blue but slightly overcast sky].*

For languages such as English, which regularly place objects and other non-subject verbal dependents after the verb (the “post-verbal domain,” Yao 2018), it seems that short constituents tend to precede longer constituents (“short before long”), as illustrated in (14). But this trend is apparently reversed for the pre-verbal domain in head-final languages like Japanese, where long constituents reportedly preferably precede short (Yamashita & Chang 2001). However, it is less clear which prediction would hold in languages which permit constituents to occur on either side of the predicate (“cross-domain NP shift,” Yao 2018). Yao (2018) investigates cross-domain NP shift for object placement in Mandarin, which varies between a post-verbal (VO) and a pre-verbal (*bǎ*-OV) option. Interestingly, this detailed study reveals no linear correlation between VO versus OV, and NP-length. Levshina (2019: 560) notes a significant effect of length only for VO languages, and most notably for clausal rather than nominal constituents. Research on diachronic syntax does consider cross-domain shift for direct objects, for example pragmatically driven object fronting in VO languages, and “heavy NP shift” in OV languages (Faarlund 2010: 205). However, for other kinds of constituent there is a general lack of research that would guide the expectations of a length effect for constituent order relative to the verb.

5.2 Method and results

In the absence of a clear hypothesis from the literature, here we present an initial exploration of length effects on pre- versus post-verbal placement of different constituents in the WOWA sample. It is important to note that overall, the leverage of the weight factor is significantly reduced in our spoken-language WOWA

data, when compared to the written-language or experimental corpora that form the basis of most previous research. Schnell & Schiborr (2022: 178) observe that in the spoken language corpora from the Multi-CAST collection (Haig & Schnell 2023), almost 90% of all NPs in the data consist of maximally three words, with the majority being two words or less. In written Universal Dependency corpora on the other hand, 36% of NPs contain four or more words (Schnell & Schiborr 2022: 178–179). Note furthermore that for WOWA, we have not included clausal constituents in our analyses, which rules out the kinds of very long tokens that figure in written language corpora (complement clauses, NPs with embedded relative clauses, etc.).

For the WOWA project, we operationalized “weight” with two measures: (i) length of the object phrase in words; and (ii) a finer-grained metric using characters, which provides a proxy for phonological weight. The data shows a strong power law-like distribution of weights, with 64% of analyzed tokens in certain roles (direct objects, Goals, Recipients and Addressees, as well as Locations, Sources, Instrumentals, Benefactives, and Comitatives) consisting of a single word and 91% of two or fewer words. Similarly, for weight in characters, 67% of these tokens contain 8 or fewer characters, and 91% contain 13 or fewer.

Two analyses were conducted both based on the finer-grained character-based weight metric. The first includes 33 data sets in WOWA, split by role; we exclude those data sets that have fewer than 8 observations in a role or display no variation in positioning. Table 4 shows the mean correlation (Pearson) between pre- and post-verbal positioning (0, 1) and weight.

Table 4: Pearson correlation of weight with position

Roles	<i>R</i> -value	SD	Observations
direct objects	+0.007	0.117	8364
goals	+0.022	0.161	4172
recipients+addressees	−0.046	0.251	1340
other obliques	+0.005	0.103	4444

All correlation coefficients hover around zero, with no substantial variation between data sets. Only a small handful of data sets have a coefficient exceeding a value of ± 0.4 in any of the roles, three of which are for Recipients/Addressees, which due to their comparative rarity unfortunately offer the least robust results in general.

The second analysis takes into consideration the basic word order of each doculect, because as Wasow (2022: 11) notes, different predictions hold for head final versus other languages. Consequently, we follow Levshina (2019) and divide the sample doculects into three groups, based on the frequency of nominal post-verbal direct objects in the corpora (the sample is not balanced across these three groups, due to the dominance of OV Iranian languages): “OV” (<25% VO), N=13; “mixed” (25–75% VO), N=16; and “VO” (>75% VO), N=4. Figure 3 shows the mean weight in characters for four role types, distinguishing pre- and post-verbal placement. For each role type, we present the findings split according to the three word-order types mentioned above (<25% VO; 25–75% VO; >75% VO).



Figure 3: Mean weight in characters of pre- and post-verbal constituents for four role types, split according to word-order type of the doculect

Figure 3 suggests a weak correlation between post-verbal placement and higher weight, for direct objects (bottom right panel) across all word order types,

but none of the individual differences reach significance. Any claim for a correlation can only therefore draw on the fact that the minimal differences are in the same direction for each language type. For other roles, no consistent pattern can be identified. Turning to word-order type, it is only the >75% VO languages that exhibit a weak but consistent association of weight and likelihood of post-verbal placement, with the strongest effect occurring for Goals in dominantly VO languages; this would merit closer investigation, but note the low absolute figures (n=15) for pre-verbal Goals in the four VO doculects.

In sum, our investigation of the effect of weight reveals only weak effects for only some roles, and some language types. This is partly attributable to the aforementioned narrow envelope of variation for weight in spoken language, where the overwhelming majority of tokens consist of only one to two words. As noted, clausal constituents were not considered in our data. Equally, we emphasize that our investigation considers weight as a factor in cross-domain shift, i.e. shift across the predicate, as opposed to relative position of constituents on the same side of the predicate, which has been the focus of most existing research (see Wasow 2022). As Yao (2018) notes, research on cross-domain weight effects is scarcely available, and their own results, like ours, reveal no clear weight effects of weight. We provisionally conclude that weight effects noted in the literature do not carry over to cross-domain word order variation in spoken language. This is definitely an area that would merit further research; see Skopeteas 2024 [this volume], on the contrasting prosodic properties that hold in the pre- and post-verbal domain respectively.

6 Summary and prospects, residual issues

In sections one to three above, we have outlined the rationale, research context, and methodologies implemented in the WOWA project. In Sections 4 and 5, we illustrate two use cases for exploring the entire database. The results for semantic role provide abundant evidence for the special role of spatial Goals in word order variability, in particular of the dominantly OV languages of Western Asia. The findings for weight, however, do not yield a simple picture, suggesting that cross-domain word-order variation (Yao 2018) requires a distinct set of explanations to those that have been proposed for same-domain word order, which focusses on the respective ordering of constituents occurring on the same side of the verb (Wasow 2022).

We hope that our research will stimulate further research in this direction, and that in the future, the hitherto neglected effects of semantic role are afforded due

consideration. In Section 5 we demonstrate that role provides the best overall predictor of post-verbal placement, with Goals outstripping any other roles by a considerable margin. We formulated our findings in the form of an implicational universal (13), which embodies a number of testable hypotheses for future work on spoken-language corpora, and also has considerable implications for understanding word order change.

Our findings also lend broad support for the concept of Transition Zone, indicating a gradual shift towards higher frequencies of verb-final constituents in the westward regions of WATZ. However, we require a more balanced and denser sample of doculects to develop a more robust framework for mapping structural variation to geospatial features, and to control for phylogenetic distance. Other issues that remain to be considered are measures of corpus-internal variation (see Craevschi 2022 for provisional findings), co-argument effects, and the role of additional morphosyntactic features such as agreement, clause type (modality, negation, subordination, etc.), and more nuanced controlling for information structure (see Hodgson et al. 2024 [this volume], Noorlander 2024a [this volume],b [this volume]).

Finally, our data point to the potential impact of register and modality (spoken versus written language) on word order. While the overwhelming majority of data in WOVA represents informal spoken language, in those data sets where data from more formal registers are available, they indicate some striking differences in word order (Nourzaei & Haig 2024 [this volume], Rasekh-Mahand et al. 2024 [this volume] and Parizadeh & Rasekh-Mahand 2024 [this volume] on Persian, Hodgson et al. 2024 [this volume] on East Armenian). These differences invite further research, but in the meantime we urge caution when comparing cross-linguistic data, and emphasize the necessity for controlling for modality and register.

7 The organization of the volume

The volume consists of 16 chapters, divided into four sections, each of which is introduced below:

- I Theoretical and methodological issues (Chapters 1–3);
- II Case studies from Iranian and Indo-Aryan languages (Chapters 4–9);
- III Case studies from the Caucasus and Black Sea (Chapters 10–13);
- IV Case studies from Semitic languages (Chapters 14–16).

7.1 Section I: Theoretical and methodological issues

Section I includes the current introductory chapter, and two further chapters. Chapter 2 by Kateryna Iefremenko investigates elements in the post-verbal domain of young adult bilingual speakers of Kurmanji and Turkish in Ankara in comparison with Turkish in Erzurum and under consideration of the sociolinguistic dichotomy between Turkish as dominant national language and Kurmanji as regional language. Although the findings on elements in the post-verbal domain in the two languages are generally in line with previous research, the results show that the Turkish dialect of Erzurum tends to have more frequent post-verbal Goals than other varieties of Turkish, which only apply post-verbal positions based on information structure and weight considerations. The higher rates of post-verbal Goals in the Erzurum dialect may plausibly reflect contact influence from neighbouring Kurmanji Kurdish dialects, which exhibit typical Iranian post-verbal Goals of motion and caused motion. In one particular construction *erdê ketin* ‘to fall on the ground’ the Goal nevertheless appears predominantly pre-verbal among bilingual Kurmanji Kurdish speakers; it may be potentially modelled on Tr. *yere düşmek* ‘to fall on the ground.’ The methodology of this study is unique in the context of the volume, in that explicitly bilingual data were analysed that were elicited from the speakers by means of video prompts.

In Chapter 3, Stavros Skopeteas investigates prosodic structure in the pre-verbal and post-verbal domain in a sample of (primarily OV) languages that includes Turkish, Georgian, Caucasian Urum, Eastern Armenian, and Persian. Skopeteas identifies three main types of OV languages, distinguished according to the nature of constraints that determine whether objects may occur in the post-verbal domain. In some languages, post-verbal objects are very restricted, and are only permitted if they are outside the focus domain of the clause, i.e. express given information, or afterthoughts (e.g. Standard Turkish). In other languages, post-verbal objects are permitted as part of broad sentence focus (e.g. Persian), while in others, even objects with narrow focus are also permitted (e.g. Georgian). In a sense, then, these three types represent increasing levels of tolerance for the integration of focal material into the post-verbal domain. The author reviews extant research on these languages and reports experimental results that illustrate the typology, and explores the interaction of prosodic and syntactic phrasing. This line of research complements the corpus-based approach of most contributions, which capture frequency patterns of linear ordering in naturalistic discourse, but leaves little space for systematic investigation of prosodic structure.

7.2 Section II: Case studies from Iranian and Indo-Aryan languages

This is the largest section in the book and contains six chapters, each of which deal with one (or a group of) Iranian or Indo-Aryan languages. In Chapter 4, Maryam Nourzaei and Geoffrey Haig present an overview of word order across three varieties of Balochi, each from areally diverse locations. The results provide further confirmation of the overall trend identified in this volume, that proximity to Mesopotamia correlates with an overall increase in post-verbal constituents; the westernmost variety of Balochi (Koroshi) exhibits both overall higher frequencies of post-verbal constituents, but also a greater range of role types permitted in this position, when compared to the two more easterly varieties. In Chapter 5, Agnes Korn presents data from two varieties of Bashkardi in southern Iran. The data stem from legacy materials recorded in the 1950's, providing a rare opportunity to explore the possibility of recent changes in the language, but also for considering micro-variation across the two varieties.

Chapter 6 (Maryam Nourzaei) illustrates the only Indo-Aryan language in the sample, Kholosi, a language island in southern Iran that has preserved aspects of Indo-Aryan morphosyntax, but has adapted in word order to conform with the post-verbal placement of spatial Goals that characterizes all of its currently neighbouring languages. In Chapter 7, Mohammad Rasekh-Mahand and co-authors provide an in-depth study of spoken Persian, comparing the recent HamBam data with the results of Frommer (1981). For the least formal registers of Persian, they report stable values over the 40-year time span with regards to most aspects of post-verbal syntax, but note a shift in register distribution in the modern data when compared to the older sample. Chapter 8 is the sole chapter based on written data, and investigates a sample of Early New Persian texts (10–13th Century CE). The texts reveal some internal variation, but an overall remarkably consistent degree of verb-finality, with little evidence for the post-verbal syntax that characterizes all spoken western Iranian languages investigated so far. These findings raise questions regarding the diachronic development of post-verbal syntax in West Iranian, but also regarding the relationship between the spoken and written languages; it is possible (and we believe plausible) that the Early New Persian texts are not representative of the spoken language of the time, any more than today's formal written Persian texts are representative of contemporary spoken Persian. In Chapter 9, Masoud Mohammadirad takes a comparative look at three varieties from the Zagros region (Gorani Gawraju; Central Kurdish Sanandaj; Southern Kurdish Bijar). The findings are suggestive of Gorani substrate effects in southernmost dialects of Central Kurdish.

7.3 Section III: Case studies from the Caucasus and Black Sea

In Chapters 10 and 11, Diana Forker investigates word order in Kartvelian and East Caucasian and Adyghe respectively. The data come from several sources, mostly outside the WOWA framework, but can be interpreted within the same framework. Role effects (Goals) are noticeable, though considerably less prevalent than in the Iranian languages and other languages of Mesopotamia. In Chapter 12, Laurentia Schreiber and Mark Janse investigate word order patterns in Romeyka in bilingual speakers under language shift to Turkish. While information structure and phrase type are the most relevant factors determining the dominant word orders in Romeyka, significant inter-speaker variation indicates the ongoing process of language shift. Chapter 13 presents original spoken-language data from East Armenian (Katherine Hodgson, Victoria Khurshudyan and Pollet Samvelian). This research adds a new perspective to the growing literature on word order in East Armenian, complementing existing research based on experimental and written-language data. The authors confirm a definiteness effect on direct object ordering, with definite direct objects showing greater word-order flexibility with respect to the verb (higher frequency of VO ordering), while indefinite objects remain overwhelmingly OV. They also identify inter-speaker variation and the effect of register. The Goals Last effect documented for most of the language of WATZ (Section 4 above) is also confirmed in these data, though in somewhat weaker magnitude than in the Iranian languages of Mesopotamia.

7.4 Section IV: Case studies from Semitic languages

This section includes three contributions on Semitic languages. In Chapter 14, Bettina Leitner describes the basic word order profile of Khuzestani Arabic, a linguistic island of Arabic in Iran, and discusses reasons for deviations from the default word order VX, such as language contact and internal change. In Chapter 15, Paul Noorlander discusses Neo-Aramaic dialects in Iran and northeastern Iraq, which include at least one dialect that has undergone a complete shift from VO to OV (Jewish Urmi). While the impetus for the shift is almost certainly language contact, Noorlander illustrates how internal factors, in particular information structure, shape the way these changes have played out. Paul Noorlander also contributes Chapter 16 on Arabic and Neo-Aramaic in Eastern Anatolia, a region of high linguistic diversity. Noteworthy findings include the variability in copula construction *s*, which contrasts with the otherwise fairly regular presence of clause-final copula elements in most of WATZ.

References

- Anonby, Christina van der Wal. 2015. *A grammar of Kumzari: A mixed Perso-Arabian language of Oman*. Leiden: Leiden University. (Doctoral dissertation).
- Asadpour, Hiwa. 2022a. *Typologizing word order variation in northwestern Iran*. Frankfurt am Main: Goethe-Universität. (Doctoral dissertation).
- Asadpour, Hiwa. 2022b. Word order in Mukri Kurdish - The case of incorporated targets. In Hiwa Asadpour & Thomas Jügel (eds.), *Word order variation: Semitic, Turkic and Indo-European languages in contact*, 63–87. Berlin: De Gruyter Mouton.
- Asadpour, Hiwa & Thomas Jügel (eds.). 2022. *Word order variation: Semitic, Turkic and Indo-European languages in contact*. Berlin: De Gruyter Mouton.
- Bağrıaçık, Metin. 2018. *Pharasiot Greek: Word order and clause structure*. Ghent: Ghent University. (Doctoral dissertation).
- Berez-Kroeker, Andrea L, Lauren Gawne, Susan Smythe Kung, Barbara F Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I Beaver, Shobhana Chelliah, Stanley Dubinsky, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.
- Bickel, Balthasar. 2017. Areas and universals. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 40–54. Cambridge: Cambridge University Press.
- Chafe, Wallace. 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.
- Comrie, Bernard. 1989. *Language universals and linguistic typology*. 2nd edn. Chicago: University of Chicago Press.
- Craevschi, Alexandru. 2022. *Historical contingency and typological tendencies in languages of Western Asia: A quantitative study of word order of non-subject constituents*. Bamberg: University of Bamberg. (MA thesis).
- Dabir Moghaddam, Mohammad. 2018. Typological approaches and dialects. In Anousha Sedighi & Pounesh Shabani-Jadidi (eds.), *The Oxford handbook of Persian linguistics*, 52–88. Oxford: Oxford University Press.
- Demir, Netice & Mahîr C. Doğan. 2021a. Zazakî (Çewlîg). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Demir, Netice & Mahîr C. Doğan. 2021b. Zazakî (Siwêreg). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in*

- word order variation. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Dogan, Mahîr. 2021a. Oghuz (Erzurum). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. (accessed on October 20, 2022). Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Dogan, Mahîr. 2021b. Oghuz (Gagauz). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Dogan, Mahîr. 2022. Vafsi (Gurchani). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Dryer, Matthew S. 1997. On the six-way word-order typology. *Studies in Language* 21(1). 69–103.
- Dryer, Matthew S. 2007. Word order. In Timothy Shopen (ed.), *Language typology and syntactic description*, vol. 1, 61–131. Cambridge: Cambridge University Press.
- Dryer, Matthew S. 2013a. Against the six-way order typology, again. *Studies in Language* 37. 267–301.
- Dryer, Matthew S. 2013b. 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. <http://wals.info/chapter/85>.
- Dryer, Matthew S. 2013c. 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. <http://wals.info/chapter/83>.
- Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63(4). 805–855. DOI: 10.2307/415719.
- Du Bois, John. 2003. Argument structure: Grammar in use. In John Du Bois, Lorraine Kumpf & William Ashby (eds.), *Preferred argument structure*, 11–60. Amsterdam: Benjamins.
- Faarlund, Jan Terje. 2010. Word order. In Silvia Luraghi & Vít Bubeník (eds.), *Continuum companion to historical linguistics*, 201–211. London/New York: Continuum.

- Faghiri, Pegah & Pollet Samvelian. 2020. Word order preferences and the effect of phrasal length in SOV languages: Evidence from sentence production in Persian. *Glossa: A journal of general linguistics* 5(1). DOI: 10.5334/GJGL.1078.
- Faghiri, Pegah, Pollet Samvelian & Barbara Hemforth. 2018. Is there a canonical order in Persian ditransitive constructions? Corpus based and experimental studies. In Agnes Korn & Andrej Malchukov (eds.), *Ditransitive constructions in a cross-linguistic perspective*, 165–185. Wiesbaden: Reichert.
- Forker, Diana. 2024a. Post-predicate elements in Adyghe. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 309–336. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266351.
- Forker, Diana. 2024b. Post-predicate elements in Kartvelian and East Caucasian. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 283–307. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266349.
- Friedman, Victor A & Brian D Joseph. 2017. Reassessing Sprachbunds: A view from the Balkans. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 55–87. Cambridge: Cambridge University Press.
- Frommer, Paul. 1981. *Post-verbal phenomena in colloquial Persian syntax*. Los Angeles: University of Southern California. (Doctoral dissertation).
- Futrell, Richard, Roger P Levy & Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language* 96(2). 371–412.
- Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2021. Typometrics from implicational to quantitative universals in word order typology. *Glossa: A journal of general linguistics* 6(1). 17.1–31.
- Haig, Geoffrey. 2015a. Verb-goal (VG) word order in Kurdish and Neo-Aramaic: Typological and areal considerations. In Geoffrey Khan & Lidia Napiorkowska (eds.), *Neo-Aramaic and its linguistic context*, 407–425. Piscataway, NJ: Gorgias Press.
- Haig, Geoffrey. 2015b. VG-word order in Kurdish and Neo-Aramaic: Typological and areal factors. *Presentation at the Cambridge Neo-Aramaic Conference, Cambridge University, 6-8th July 2011*. Cambridge.
- Haig, Geoffrey. 2017. Western Asia: East Anatolia as a transition zone. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 396–423. Cambridge: Cambridge University Press.

- Haig, Geoffrey. 2022a. Balochi (Turkmenistan). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Haig, Geoffrey. 2022b. Kumzari (Musandam). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Haig, Geoffrey. 2022c. Kurdish (Northern, Muş). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Haig, Geoffrey. 2022d. Post-predicate constituents in Kurdish. In Yaron Matras, Geoffrey Haig & Ergin Öpengin (eds.), *Structural and typological variation in the dialects of Kurdish*, 335–377. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-78837-7_8.
- Haig, Geoffrey & Geoffrey Khan. 2019. Introduction. In Geoffrey Haig & Geoffrey Khan (eds.), *The languages and linguistics of Western Asia: An areal perspective*, 1–29. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110421682-001.
- Haig, Geoffrey, Paul M. Noorlander & Nils N. Schiborr. In press. Which word order features are stable in a contact setting? Corpus-based evidence from the Western Asian transition zone. In Jeroen Darquennes, Joe Salmons & Wim Vandebussche (eds.), *Language contact: An international handbook*, vol. 2. Berlin: De Gruyter Mouton.
- Haig, Geoffrey & Mohammad Rasekh-Mahand. 2022. *HamBam: The Hamedan-Bamberg corpus of contemporary spoken Persian*. multicast.aspra.uni-bamberg.de/resources/hambam/ (10 July, 2023).
- Haig, Geoffrey & Nils N. Schiborr. In review. Goals last in grammar and discourse.
- Haig, Geoffrey & Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618. DOI: 10.1353/lan.2016.0049.
- Haig, Geoffrey & Stefan Schnell. 2023. *Multi-CAST: Multilingual corpus of annotated spoken texts. (Version 2311)*. Bamberg. multicast.aspra.uni-bamberg.de.
- Haig, Geoffrey, Stefan Schnell & Nils N. Schiborr. 2021. Universals of reference in discourse and grammar: Evidence from the multi-CAST collection of spoken corpora. In Geoffrey Haig, Stefan Schnell & Frank Seifart (eds.), *Doing corpus-*

- based typology with spoken language data: State of the art*, 141–177. Honolulu: University of Hawai'i Press.
- Haig, Geoffrey, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr. 2022. *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Haig, Geoffrey & Hannah Thiele. 2014. Post-predicate Goals in Northern Kurdish and neighbouring languages: A pilot study in quantitative areal linguistics languages. Paper presented at the *Second International Conference on Variation and Change in Kurdish*, mardin artuklu university, turkey, 8-9 October 2014.
- Harris, Alice C. & Lyle Campbell. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.
- Haspelmath, Martin. 2015. Ditransitive constructions. *Annual Review of Linguistics* 1(1). 19–41. DOI: [10.1146/annurev-linguist-030514-125204](https://doi.org/10.1146/annurev-linguist-030514-125204).
- Hawkins, John A. 2008. An asymmetry between VO and OV languages: The ordering of obliques. In Greville Corbett & Michael Noonan (eds.), *Case and grammatical relations: Studies in honor of Bernard Comrie*, 167–190. Amsterdam: Benjamins. DOI: doi.org/10.1075/tsl.81.08ana.
- Hodgson, Katherine, Victoria Khurshudyan & Pollet Samvelian. 2024. Post-predicate arguments in Modern Eastern Armenian. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 375–410. Berlin: Language Science Press. DOI: [10.5281/zenodo.14266355](https://doi.org/10.5281/zenodo.14266355).
- Iefremenko, Kateryna. 2021a. Kurdish (Northern, Ankara). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. (accessed on October 20, 2022). Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa.
- Iefremenko, Kateryna. 2021b. Oghuz (Ankara). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/#turkic.
- Izadi, Elham. 2022. Persian (New). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.

- Izadifar, Raheleh. 2022. Tāti (Hazārrudi). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Jahani, Carina. 2018. Post-verbal arguments in Balochi. Paper presented at *the International Conference on Anatolia-Caucasus-Iran: Ethnic and linguistic contacts (ACIC)*, Yerevan University, 10–12 May.
- Jing, Yingqi, Paul Widmer & Balthasar Bickel. 2021. Word order variation is partially constrained by syntactic complexity. *Cognitive Science* 45(11). 1–25. DOI: 10.1111/cogs.13056.
- Khan, Geoffrey. 2008a. *The Jewish Neo-Aramaic dialect of Urmi*. Piscataway, NJ: Gorgias Press.
- Khan, Geoffrey. 2008b. *The Neo-Aramaic dialect of Barwar*. Leiden: Brill. DOI: 10.1163/ej.9789004167650.i-2198.
- Korn, Agnes. 2024. Notes on word order in Bashkardi. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 155–173. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266339.
- Korn, Agnes & Ilya Gershevitch. 2023a. Bashkardi (Northern): Recordings by Ilya Gershevitch (1914–2001), transcription and analysis by Agnes Korn. In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Korn, Agnes & Ilya Gershevitch. 2023b. Bashkardi (Southern): Recordings by Ilya Gershevitch (1914–2001), transcription and analysis by Agnes Korn. In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Lazard, Gilbert. 1957. *Grammaire du Persan contemporain*. Paris: Klincksieck.
- Leitner, Bettina. 2021. Arabic (Khuzestan). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg. <http://multicast.aspra.uni-bamberg.de/resources/wowa/> (6 March, 2023).

- 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.
- Levshina, Natalia, Savithry Namboodiripad, Marc Allasonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova & Natalia Stojnova. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825–883. DOI: 10.1515/ling-2021-0098.
- MacKenzie, David N. 1962. *Kurdish dialect studies Vol. II*. London/New York: Oxford University Press.
- Mettouchi, Amina & Martine Vanhove. 2021. Prosodic segmentation and cross-linguistic comparison in corpfraos and cortypo: corpus-driven and corpus-based approaches. In Geoffrey Haig, Stefan Schnell & Frank Seifart (eds.), *Doing corpus-based typology with spoken language corpora: State of the art (language documentation & conservation special publication 25)*, 59–113. Honolulu, HI: University of Hawai'i Press.
- Mithun, Marianne. 1992. Is basic word order universal? Grounding and coherence in discourse. In Doris Payne (ed.), *The pragmatics of word-order flexibility*, 15–61. Amsterdam: Benjamins.
- Mohammadirad, Masoud. 2022a. Gorani (Gawraǰū). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: Bamberg University Press. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Mohammadirad, Masoud. 2022b. Kurdish (Central, Sanandaj). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Mohammadirad, Masoud. 2022c. Kurdish (Southern, Bijar). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: Bamberg University Press. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Mohammadirad, Masoud. 2024. Zagros region: The Kurdish-Gorani continuum. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian*

- Transition Zone: A corpus-based approach to areal typology*, 245–279. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266347.
- Molin, Dorota. 2022. NE Neo-Aramaic (Jewish Dohok). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa (5 July, 2023).
- Neocleous, Nicolaos. 2020. *Word order and information structure in Romeyka: A syntax and semantics interface account of order in a minimalist system*. Cambridge: University of Cambridge. (Doctoral dissertation).
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th language Resources and Evaluation Conference*, 4034–4043. Marseille: European Language Resources Association.
- Noorlander, Paul M. 2022a. Arabic (Jewish, Baghdad). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Noorlander, Paul M. 2022b. NE Neo-Aramaic (Jewish Sanandaj). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Noorlander, Paul M. 2024a. Arabic and Aramaic in eastern Anatolia. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 471–514. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266361.
- Noorlander, Paul M. 2024b. Neo-Aramaic in Iran and northeastern Iraq. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 431–469. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266359.
- Nourzaei, Maryam. 2021a. Balochi (Coastal). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order vari-*

- ation. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Nourzaei, Maryam. 2021b. Balochi (Koroshi). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Nourzaei, Maryam. 2022. Kholosi (Kholos). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Nourzaei, Maryam. 2024. Kholosi. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 175–195. Berlin: Language Science Press. DOI: [10.5281/zenodo.14266341](https://doi.org/10.5281/zenodo.14266341).
- Nourzaei, Maryam & Geoffrey Haig. 2024. Balochi: A cross-dialect investigation of post-verbal elements. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 121–154. Berlin: Language Science Press. DOI: [10.5281/zenodo.14266337](https://doi.org/10.5281/zenodo.14266337).
- Parizadeh, Mehdi. 2022. Persian (Early new). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Parizadeh, Mehdi & Mohammad Rasekh-Mahand. 2024. Post-predicate elements in Early New Persian (10–13th Century CE). In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*, 231–244. Berlin: Language Science Press. DOI: [10.5281/zenodo.14266345](https://doi.org/10.5281/zenodo.14266345).
- Petersen, Jan Heegård. 2015. Kalasha texts with introductory grammar. *Acta Linguistica Hafniensia* 47(S1). 1–275. DOI: [10.1080/03740463.2015.1069049](https://doi.org/10.1080/03740463.2015.1069049).
- Rasekh-Mahand, Mohammad, Elham Izadi, Mehdi Parizadeh, Geoffrey Haig & Nils Schiborr. 2024. Post-predicate elements in modern colloquial Persian: A multifactorial analysis. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in*

- the Western Asian Transition Zone: A corpus-based approach to areal typology*, 197–230. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266343.
- Robbeets, Martine. 2017. The Transeurasian languages. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 586–626. Cambridge: Cambridge University Press.
- Schnell, Stefan, Geoffrey Haig, Nils N. Schiborr & Maria Vollmer. 2023. Are referent introductions sensitive to forward planning in discourse? Evidence from Multi-CAST. In Simone Mattioli & Alessandra Barotto (eds.), *Discourse phenomena in typological perspective*, 231–268. Amsterdam: Benjamins.
- Schnell, Stefan, Geoffrey Haig & Frank Seifart. 2021. The role of language documentation in corpus-based typology. In Geoffrey Haig & Stefan Schnell (eds.), *Doing corpus-based 38 typology with spoken language data: State of the art*, 1–28. Honolulu: University of Hawai'i Press.
- Schnell, Stefan & Nils N. Schiborr. 2022. Crosslinguistic corpus studies in linguistic typology. *Annual Review of Linguistics* 8(1). 171–191. DOI: 10.1146/annurev-linguistics-031120-104629.
- Schreiber, Laurentia. 2021a. Oghuz (Qashqai). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahir Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Schreiber, Laurentia. 2021b. Pontic Greek (Romeyka). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahir Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Schreiber, Laurentia, Mortaza Taheri-Ardali, Geoffrey Haig & Erik Anonby. 2021. Contact-induced change in the morphosyntax of Turkic in Boldaji, Chahar Mahal va Bakhtiari province, Iran. *Turkic Languages* 25(2). 210–242. DOI: 10.13173/TL.25.2.210.
- Seifart, Frank, Ludger Paschen & Matthew Stave (eds.). 2022. *Language documentation reference corpus (DoReCo)*. Berlin/Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage. DOI: 10.34847/nkl.7cbfq779.
- Siewierska, Anna. 1988. *Word order rules*. London: Croom Helm.
- Skopeteas, Stavros. 2024. Post-predicate prosody in OV languages. In Geoffrey Haig, Mohammad Rasekh-Mahand, Donald Stilo, Laurentia Schreiber & Nils N. Schiborr (eds.), *Post-predicate elements in the Western Asian Transition Zone: A*

- corpus-based approach to areal typology*, 87–117. Berlin: Language Science Press. DOI: 10.5281/zenodo.14266335.
- Song, Jae Jung. 2018. *Linguistic typology*. Oxford: Oxford University Press.
- Stilo, Donald. 2005. Iranian as a buffer zone between the universal typology of Turkic and Semitic. In Éva Csató, Bo Isaksson & Carina Jahani (eds.), *Linguistic convergence and areal diffusion. Case studies from Iranian, Semitic and Turkic*, 35–63. London/New York: Routledge Curzon.
- Stilo, Donald. 2006. Circumpositions as an areal response: The case study of the Iranian zone. In Lars Johanson & Christiane Bulut (eds.), *Turkic-Iranian contact areas: Historical and linguistic aspects*, 310–333. Wiesbaden: Harrassowitz.
- Stilo, Donald. 2009. Circumpositions as an areal response: The case study of the Iranian zone. *Turkic Languages* 13. 3–33.
- Stilo, Donald. 2012. Intersection zones, overlapping isoglosses, and “Fade-out/Fade-in” phenomena in Central Iran. In Mohammad R. Ghanoonparvar & Behrad Aghaei (eds.), *Iranian languages and culture: Essays in honor of Gernot Ludwig Windfuhr*, 134–155. Costa Mesa, CA: Mazda Publisher.
- Stilo, Donald. 2018a. Investigating shared features in the Araxes-Iran linguistic area and its subareas. In Christiane Bulut (ed.), *Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery*, 427–452. Wiesbaden: Harrassowitz.
- Stilo, Donald. 2018b. Preverbal and postverbal peripheral arguments in the Araxes-Iran linguistic area. Paper presented at the *Conference Anatolia-Caucasus-Iran: Ethnic and Linguistic Contacts* 10-12 May 2018. Yerevan University, Yerevan, Armenia.
- Stilo, Donald. 2021a. Oghuz (Tabriz). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Stilo, Donald. 2021b. Talyshi (Lerik). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Stilo, Donald. 2022a. Kurdish (Northern, Lachin). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.

- Stilo, Donald. 2022b. NE Neo-Aramaic (Christian) Barwar. In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: Bamberg University. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Stilo, Donald & Geoffrey Haig. 2022. Mazandarani (Kordxeyl). In Geoffrey Haig, Donald Stilo, Nils N. Schiborr & Mahîr Dogan (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/.
- Stilo, Donald & René Lacroix. 2021. Laz (Arhavi). In Geoffrey Haig, Donald Stilo, Mahîr C. Doğan & Nils N. Schiborr (eds.), *WOWA – Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/resources/wowa/ (27 July, 2023).
- Szeto, Pui Yiu & Chingduang Yurayong. 2021. Sinitic as a typological sandwich: Revisiting the notions of altaicization and taicization. *Linguistic Typology* 25(3). 551–599.
- Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77–94. DOI: 10.1515/LITY.2009.004.
- Wasow, Thomas. 2022. Factors influencing word ordering. In Hiwa Asadpour & Thomas Jügel (eds.), *Word order variation: Semitic, Turkic and Indo-European languages in contact*, 1–14. Berlin: De Gruyter Mouton.
- Yamashita, Hiroko & Franklin Chang. 2001. “Long before short” preference in the production of a head-final language. *Cognition* 81(2). B45–B55. DOI: 10.1016/S0010-0277(01)00121-4.
- Yao, Yao. 2018. NP weight effects in word order variation in Mandarin Chinese. *Lingua Sinica* 4(5). 1–31. DOI: 10.1186/s40655-018-0037-8.

Appendix A Data sources and raw figures

Table 5: Data sources: 35 doculects in WOWA (April 2024). Legend for abbreviations: OV = object-verb word order; VO = verb-object word order; NENA = North Eastern Neo-Aramaic. “Unpubl.” indicates data-sets which are fully annotated, but due to accessibility issues cannot be published online.

doculect	affiliation	source
<i>Oghuz (Ankara)</i>	● Turkic	Iefremenko 2021b
<i>Oghuz (Bayat)</i>	● Turkic	Unpubl.
<i>Oghuz (Erzurum)</i>	● Turkic	Dogan 2021a
<i>Oghuz (Gagauz)</i>	● Turkic	Dogan 2021b
<i>Oghuz (Qashqai)</i>	● Turkic	Schreiber 2021a
<i>Oghuz (Tabriz)</i>	● Turkic	Stilo 2021a
<i>Balochi (Coastal)</i>	◆ Iranian, western	Nourzaei 2021a
<i>Balochi (Koroshi)</i>	◆ Iranian, western	Nourzaei 2021b
<i>Balochi (Turkmen)</i>	◆ Iranian, western	Haig 2022a
<i>Bashkardi (Northern)</i>	◆ Iranian, western	Korn & Gershevitch 2023a
<i>Bashkardi (Southern)</i>	◆ Iranian, western	Korn & Gershevitch 2023b
<i>Gorani (Gawraǰū)</i>	◆ Iranian, western	Mohammadirad 2022a
<i>Kumzari (Musandam)</i>	◆ Iranian, western	Haig 2022b
<i>Kurdish (Central, Sanandaj)</i>	◆ Iranian, western	Mohammadirad 2022b
<i>Kurdish (Northern, Ankara)</i>	◆ Iranian, western	Iefremenko 2021a
<i>Kurdish (Northern, Lachin)</i>	◆ Iranian, western	Stilo 2022a
<i>Kurdish (Northern, Muš)</i>	◆ Iranian, western	Haig 2022c
<i>Kurdish (Southern, Bijar)</i>	◆ Iranian, western	Mohammadirad 2022c
<i>Mazandarani (Kordxeyl)</i>	◆ Iranian, western	Stilo & Haig 2022
<i>Persian (New)</i>	◆ Iranian, western	Izadi 2022
<i>Persian (New, Early Classical)</i>	◆ Iranian, western	Parizadeh 2022
<i>Talyshi (Lerik)</i>	◆ Iranian, western	Stilo 2021b
<i>Tat (Daykušču)</i>	◆ Iranian, western	Unpubl.
<i>Tati (Hazārudi)</i>	◆ Iranian, western	Izadifar 2022
<i>Vafsi (Gurchani)</i>	◆ Iranian, western	Dogan 2022
<i>Zazakî (Çewlig)</i>	◆ Iranian, western	Demir & Doğan 2021a
<i>Zazakî (Siwêreg)</i>	◆ Iranian, western	Demir & Doğan 2021b
<i>NENA (Christian, Barwar)</i>	▲ West Semitic	Stilo 2022b
<i>NENA (Jewish, Dohok)</i>	▲ West Semitic	Molin 2022
<i>NENA (Jewish, Sanandaj)</i>	▲ West Semitic	Noorlander 2022b
<i>Arabic (Jewish, Baghdad)</i>	▲ West Semitic	Noorlander 2022a
<i>Arabic (Khuzestan)</i>	▲ West Semitic	Leitner 2021
<i>Kholosi (Kholos)</i>	✱ Indo-Aryan	Nourzaei 2022
<i>Laz (Arhavi)</i>	◆ Kartvelian	Stilo & Lacroix 2021
<i>Pontic Greek (Romeyka)</i>	◆ Hellenic	Schreiber 2021b

1 Post-predicate elements in the Western Asian Transition Zone

Table 6: Raw figures for the WOWA data sets, corpus size and mean token weights in words and characters

doculect	texts	words	tokens	token weight valid in words		token weight in characters	
				mean	SD	mean	SD
<i>Oghuz (Ankara)</i>	28	4145	587	1.42	0.68	8.80	5.18
<i>Oghuz (Bayat)</i>	1	3037	835	1.46	0.72	7.62	4.17
<i>Oghuz (Erzurum)</i>	3	3860	636	1.35	0.58	7.63	3.77
<i>Oghuz (Gagauz)</i>	2	5220	594	1.39	0.65	7.63	4.12
<i>Oghuz (Qashqai)</i>	5	2915	557	1.52	0.82	7.84	4.56
<i>Oghuz (Tabriz)</i>	13	3468	851	1.47	0.71	8.37	4.88
<i>Balochi (Coastal)</i>	3	6768	1535	1.42	0.60	8.11	4.79
<i>Balochi (Koroshi)</i>	2	3083	573	1.53	0.73	9.11	4.63
<i>Balochi (Turkmen)</i>	4	4323	580	1.60	0.84	8.25	4.94
<i>Bashkardi (Southern)</i>	5	947	234	1.35	0.63	6.85	3.37
<i>Bashkardi (Northern)</i>	6	2744	596				
<i>Gorani (Gawraju)</i>	7	8782	1015	1.35	0.64	7.21	4.12
<i>Kumzari (Musandam)</i>	2	4496	592	1.25	0.58	5.49	3.04
<i>Kurdish (Central, Sanandaj)</i>	11	8502	1180	1.37	0.62	7.48	3.91
<i>Kurdish (Northern, Ankara)</i>	30	4728	507	1.45	0.60	8.03	4.31
<i>Kurdish (Northern, Lachin)</i>	28	3714	773	1.84	0.76	7.37	4.25
<i>Kurdish (Northern, Mus)</i>	2	2711	693	1.47	0.64	4.84	1.99
<i>Kurdish (Southern, Bijar)</i>	8	7251	1150	1.45	0.72	7.53	4.61
<i>Mazandarani (Kordxeyl)</i>	7	3193	676	1.56	0.70	7.30	4.16
<i>Persian (New)</i>	30	12564	1628	1.65	0.84	9.71	6.02
<i>Persian (New, Early Classical)</i>	3	6751	1278	1.59	0.86	9.40	6.67
<i>Talyshi (Lerik)</i>	3	2872	650	1.76	0.73	7.16	3.80
<i>Tat (Daykuseu)</i>	1	1316	320	1.38	0.54	7.38	3.52
<i>Tati (Hazarrudi)</i>	8	4068	665	1.37	0.68	7.00	4.08
<i>Vafsi (Gurchani)</i>	10	4751	733	1.56	0.76	7.89	3.75
<i>Zazaki (Cewlig)</i>	1	2444	410	1.43	0.66	5.89	3.28
<i>Zazaki (Siwereg)</i>	1	1972	352	1.39	0.59	5.99	3.19
<i>NENA (Christian, Barwar)</i>	5	3517	963	1.38	0.66	7.83	4.68
<i>NENA (Jewish, Dohok)</i>	11	3295	514	1.26	0.54	6.85	3.56
<i>NENA (Jewish, Sanandaj)</i>	4	7166	1184	1.25	0.52	6.74	3.05
<i>Arabic (Jewish, Baghdad)</i>	4	3057	490	1.39	0.68	9.01	5.14
<i>Arabic (Khuzestan)</i>	6	6391	546	1.33	0.65	7.90	3.94
<i>Kholosi (Kholos)</i>	2	3171	516	1.54	0.77	8.72	4.89
<i>Laz (Arhavi)</i>	11	1389	400	1.22	0.51	7.72	4.43
<i>Pontic Greek (Romeyka)</i>	5	2946	501	1.66	0.72	8.37	3.76

Table 7: Raw figures of the WOVA data sets, rates of post-verbal placement of nominal direct objects and goals

doculect	nominal direct objects			nominal goals		
	n(post)	n(all)	%(post)	n(post)	n(all)	%(post)
<i>Oghuz (Ankara)</i>	2	88	2	9	123	7
<i>Oghuz (Bayat)</i>	10	283	4	75	117	64
<i>Oghuz (Erzurum)</i>	16	229	7	46	120	38
<i>Oghuz (Gagauz)</i>	78	154	51	64	88	73
<i>Oghuz (Qashqai)</i>	12	147	8	58	82	71
<i>Oghuz (Tabriz)</i>	2	219	1	88	117	75
<i>Balochi (Coastal)</i>	23	338	7	71	112	63
<i>Balochi (Koroshi)</i>	4	182	2	77	86	90
<i>Balochi (Turkmen)</i>	3	192	2	20	42	48
<i>Bashkardi (Southern)</i>	8	73	11	41	51	80
<i>Bashkardi (Northern)</i>	50	182	27	58	92	63
<i>Gorani (Gawraju)</i>	13	275	5	233	243	96
<i>Kumzari (Musandam)</i>	8	115	7	83	83	100
<i>Kurdish (Central, Sanandaj)</i>	3	295	1	267	283	94
<i>Kurdish (Northern, Ankara)</i>	0	81	0	70	119	59
<i>Kurdish (Northern, Lachin)</i>	3	197	2	90	111	81
<i>Kurdish (Northern, Muş)</i>	6	217	3	107	120	89
<i>Kurdish (Southern, Bijar)</i>	7	298	2	272	281	97
<i>Mazandarani (Kordxeyl)</i>	8	319	3	68	108	63
<i>Persian (New)</i>	19	377	5	218	258	84
<i>Persian (New, Early Classical)</i>	4	257	2	1	21	5
<i>Talyshi (Lerik)</i>	4	164	2	73	102	72
<i>Tat (Daykusecu)</i>	15	100	15	35	45	78
<i>Tati (Hazarrudi)</i>	4	153	3	111	121	92
<i>Vafsi (Gurchani)</i>	4	257	2	146	166	88
<i>Zazaki (Cewlig)</i>	4	85	5	90	99	91
<i>Zazaki (Siwereg)</i>	4	86	5	46	46	100
<i>NENA (Christian, Barwar)</i>	262	315	83	105	109	96
<i>NENA (Jewish, Dohok)</i>	188	210	90	105	106	99
<i>NENA (Jewish, Sanandaj)</i>	18	331	5	171	185	92
<i>Arabic (Jewish, Baghdad)</i>	159	164	97	77	77	100
<i>Arabic (Khuzestan)</i>	267	308	87	77	81	95
<i>Kholosi (Kholos)</i>	2	138	1	34	55	62
<i>Laz (Arhavi)</i>	4	128	3	2	54	4
<i>Pontic Greek (Romeyka)</i>	116	175	66	62	78	7

1 Post-predicate elements in the Western Asian Transition Zone

Table 8: Raw figures of the WOVA data sets, rates of post-verbal placement of pronominal direct objects and goals

doculect	pronominal direct objects			pronominal goals		
	n(post)	n(all)	%(post)	n(post)	n(all)	%(post)
<i>Oghuz (Ankara)</i>	1	14	7	2	19	11
<i>Oghuz (Bayat)</i>	6	70	9	4	8	50
<i>Oghuz (Erzurum)</i>	4	54	7	1	11	9
<i>Oghuz (Gagauz)</i>	26	64	41	7	11	64
<i>Oghuz (Qashqai)</i>	1	28	4	6	11	55
<i>Oghuz (Tabriz)</i>	6	59	10	11	16	69
<i>Balochi (Coastal)</i>	27	99	27	1	2	50
<i>Balochi (Koroshi)</i>	0	18	0	0	1	0
<i>Balochi (Turkmen)</i>	2	55	4	0	5	0
<i>Bashkardi (Southern)</i>	0	9	0	0	1	0
<i>Bashkardi (Northern)</i>	2	23	9	1	5	20
<i>Gorani (Gawraju)</i>	0	32	0	3	4	75
<i>Kumzari (Musandam)</i>	52	81	64	40	40	100
<i>Kurdish (Central, Sanandaj)</i>	0	24	0	11	13	85
<i>Kurdish (Northern, Ankara)</i>	1	11	9	5	9	56
<i>Kurdish (Northern, Lachin)</i>	0	34	0	3	5	60
<i>Kurdish (Northern, Muš)</i>	2	41	5	4	12	33
<i>Kurdish (Southern, Bijar)</i>	0	45	0	2	2	100
<i>Mazandarani (Kordxeyl)</i>	6	62	10	8	13	62
<i>Persian (New)</i>	1	63	2	1	3	33
<i>Persian (New, Early Classical)</i>	0	63	0	0	4	0
<i>Talyshi (Lerik)</i>	2	73	3	8	14	57
<i>Tat (Daykuseu)</i>	7	17	41	4	4	100
<i>Tati (Hazarrudi)</i>	4	60	7	13	17	76
<i>Vafsi (Gurchani)</i>	3	46	7	2	2	100
<i>Zazaki (Cewlig)</i>	4	41	10	11	13	85
<i>Zazaki (Siwereg)</i>	3	30	10	3	7	43
<i>NENA (Christian, Barwar)</i>	15	44	34	14	17	82
<i>NENA (Jewish, Dohok)</i>	21	31	68	6	6	100
<i>NENA (Jewish, Sanandaj)</i>	1	48	2	17	21	81
<i>Arabic (Jewish, Baghdad)</i>	13	18	72	0	2	0
<i>Arabic (Khuzestan)</i>	5	10	50	0	1	0
<i>Kholosi (Kholos)</i>	2	16	12	1	1	100
<i>Laz (Arhavi)</i>	0	35	0	0	5	0

Table 9: Raw figures for the WOWA data sets, rates of post-verbal placement of nominal and pronominal addressees/recipients and various other obliques (locations, sources, instruments, benefactives, comitatives)

doculect	recipients/addressees			other obliques		
	n(post)	n(all)	%(post)	n(post)	n(all)	%(post)
<i>Oghuz (Ankara)</i>	3	16	19	13	199	7
<i>Oghuz (Bayat)</i>	21	49	43	20	152	13
<i>Oghuz (Erzurum)</i>	4	52	8	10	116	9
<i>Oghuz (Gagauz)</i>	7	15	47	43	114	38
<i>Oghuz (Qashqai)</i>	7	8	88	13	81	16
<i>Oghuz (Tabriz)</i>	11	53	21	24	190	13
<i>Balochi (Coastal)</i>	31	112	28	25	138	18
<i>Balochi (Koroshi)</i>	10	24	42	17	64	27
<i>Balochi (Turkmen)</i>	3	21	14	7	110	6
<i>Bashkardi (Southern)</i>	3	8	38	5	14	36
<i>Bashkardi (Northern)</i>	23	41	56	15	49	31
<i>Gorani (Gawraju)</i>	27	33	82	37	102	36
<i>Kumzari (Musandam)</i>	93	96	97	61	70	87
<i>Kurdish (Central, Sanandaj)</i>	21	22	95	44	200	22
<i>Kurdish (Northern, Ankara)</i>	2	15	13	12	209	6
<i>Kurdish (Northern, Lachin)</i>	22	55	40	9	244	4
<i>Kurdish (Northern, Muš)</i>	16	45	36	7	139	5
<i>Kurdish (Southern, Bijar)</i>	11	41	27	17	148	11
<i>Mazandarani (Kordxeyl)</i>	6	50	12	9	119	8
<i>Persian (New)</i>	17	67	25	51	262	19
<i>Persian (New, Early Classical)</i>	0	41	0	16	176	9
<i>Talyshi (Lerik)</i>	23	51	45	37	142	26
<i>Tat (Daykuseu)</i>	12	35	34	17	57	30
<i>Tati (Hazarrudi)</i>	9	17	53	33	212	16
<i>Vafsi (Gurchani)</i>	14	37	38	8	72	11
<i>Zazaki (Cewlig)</i>	4	27	15	0	49	0
<i>Zazaki (Siwereg)</i>	15	29	52	4	49	8
<i>NENA (Christian, Barwar)</i>	13	13	100	148	200	74
<i>NENA (Jewish, Dohok)</i>	41	42	98	50	54	93
<i>NENA (Jewish, Sanandaj)</i>	58	72	81	57	141	40
<i>Arabic (Jewish, Baghdad)</i>	11	11	100	70	82	85
<i>Arabic (Khuzestan)</i>	5	5	100	32	34	94
<i>Kholosi (Kholos)</i>	7	18	39	1	84	1
<i>Laz (Arhavi)</i>	2	36	6	1	87	1
<i>Pontic Greek (Romeyka)</i>	3	5	60	43	114	38