The
Word Order in
Western Asia
Corpus

README

16 July 2025
(doc v1.0)

Introduction

WOWA (*Word Order in Western Asia*) is an open-access collection of transcribed and annotated spoken texts from 41 languages (or more accurately, doculects), spoken across a region loosely referred to here as Western Asia. Most texts are spontaneous (i.e. unscripted) narrative monologues such as oral history, traditional tales, etc. WOWA was funded by the Alexander-von-Humboldt Stiftung (grant number 1135327-IRN-IP), awarded to Geoffrey Haig (Bamberg) and Mohammad Rasekh-Mahand (Hamedan), 2019–2023. The archive was designed and implemented by N. Schiborr. The main results of the project have been published in Haig et al. (2024a).

Citation

The WOWA collection can be cited as follows:

> Haig, Geoffrey & Stilo, Donald & Doğan, Mahîr C. & Schiborr, N. (eds.). 2024. *WOWA — Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. (DOI: 10.48564/unibafd-gyws0-g4218)

Each data set within the collection can furthermore be treated as an individually citable resource. Please refer to the citation file accompanying each data set for a suggestion on how to do so.

Research context and outcomes

WOWA was set up specifically to investigate areal effects in word order patterns across regions characterized by the intersection of languages with diametrically opposed word order typologies, also known as transition zones. Linguistically, the project focuses on the post-verbal domain and the kinds of arguments that are systematically associated with it, which has proved to be an important locus of areal influence. The project also contributes to the documentation of the many endangered (in some cases extinct) languages of Western Asia.

See Haig et al. (2024b) for further explanations of the theory and methodology, as well as a number of case studies on semantic role and constituent weight. See further Noorlander et al. (2025) for a study of word order change in Semitic, and Haig et al. (2025) for an examination of the relative stability of typological features in contact settings.

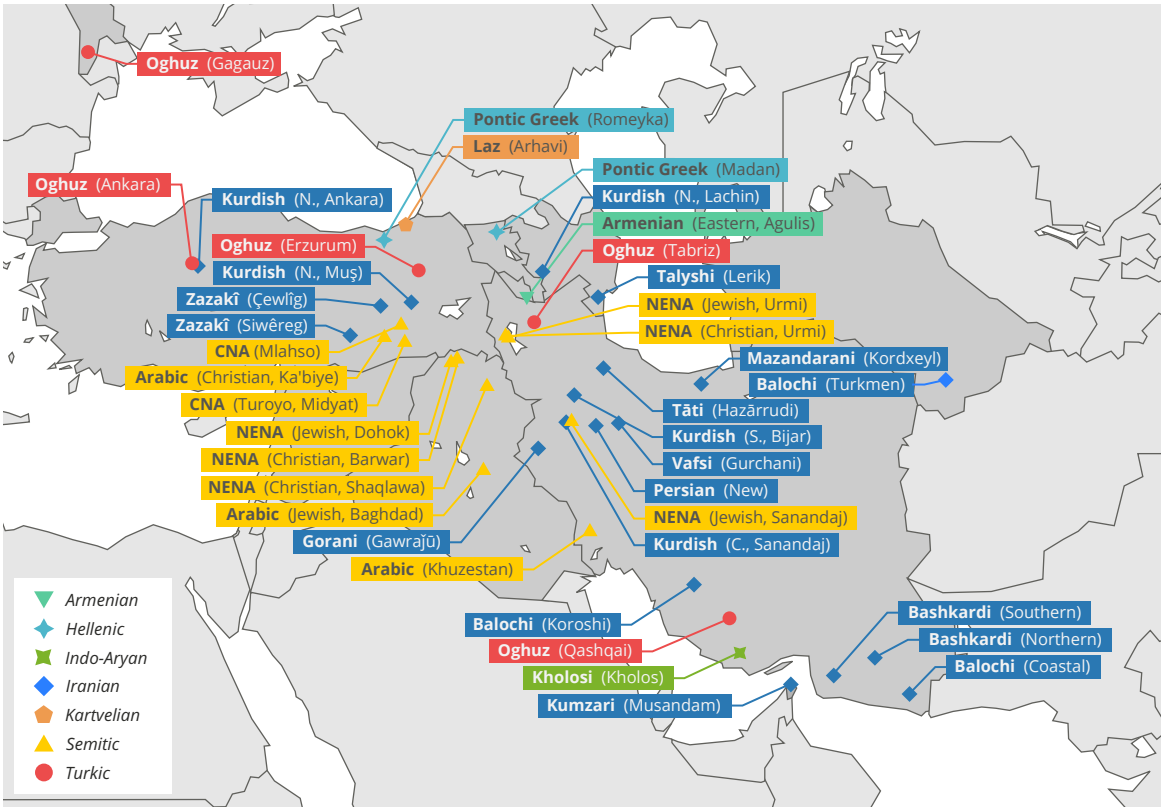


Figure 1 The WOWA doculects.

Data sources

The primary data are transcriptions of natural, unscripted spoken texts, generally monologues, such as traditional folk tales or personal histories. The data have been annotated according to a unified scheme (see further below). The sources of the texts are varied: Some stem from published legacy materials, others from recent language documentation initiatives, others from dialect surveys undertaken by various institutions. Transcriptions follow the original sources; any modifications are noted in the metadata. Most texts are accompanied by a translation into English, and many also include morphological glossing, are accompanied by sound files (WAV), or are linked to additional resources.

The languages selected are generally under-researched, non-standardized minority languages, which reflect the long-term linguistic diversity of the region more faithfully than the currently dominant written official languages (Turkish, Arabic, and Persian). Representatives of the following high-level linguistic groups are included in the data: Armenian (Indo-European), Hellenic (Indo-European), Indo-Aryan (Indo-European), Iranian (Indo-European), Kartvelian, Semitic, and Turkic. The data were contributed and annotated by a consortium of experts:

> Assaf Bar-Moshe, Alexandru Craevschi, Netîce Demir, Mahîr Doğan, Sohrab Dolatkhah, Ilya Gershevitch, Geoffrey Haig, Katherine Hodgson, Kateryna Iefremenko, Elham Izadi, Raheleh Izadifar, Agnes Korn, René Lacroix, Bettina Leitner, Masoud Mohammadirad, Dorota Molin, Paul Noorlander, Maryam Nourzaei, Mehdi Parizadeh, Laurentia Schreiber, and Donald Stilo.

Annotations and coding procedures

WOWA was designed to address word order phenomena within the languages of the Western Asian Transition Zone, specifically the position of non-subject arguments relative to the predicate. The category of non-subject arguments include direct objects and various other kinds of verbal dependents including recipients, instruments, locations, and goals. Note that only tokens of nominal constituents are coded; clausal constituents (complement clauses) are not considered: The basic unit for analysis is therefore referential, non-subject nominal constituents. Each token of such a constituent is identified in the corpus and annotated according to a number of criteria, including animacy, form class (pronominal, nominal, etc.), semantic role, flagging, etc. The detailed annotation guidelines are included in this archive. A more complete description of the research context, corpus design, and illustrative examples can be found in Haig et al. (2024b).

Archive contents

Each of the 41 data set in WOWA is an individually citable resource that includes minimally source texts (PDF, optionally also WAV audio recordings), original annotations (XLS spreadsheets and TSV), a metadata sheet (PDF), and a citation file (TXT).

The files in each data sets are named using a schema *x_y_z*, where *x* is a language's high-level affiliation (e.g. *semi* for Semitic), *y* is its mid-level affiliation (e.g. *arabic*), and *z* its low-level affiliation or doculect (e.g. *khuzestan*). The archive as a whole is split into into three parts, each wrapped into separate ZIP files:

- (1*) **wowa__annotations.zip**: annotation and documentation files
  - (1*) *annotations* folder
    - (41*) *x_y_z* folders, one for each data set
      - (1*) *wowa_x_y_z.xls*: spreadsheet with original annotations
      - (1*) *wowa_x_y_z.tsv*: TSV version of the spreadsheet
      - (1*) *wowa_x_y_z__metadata.pdf*: metadata sheet
      - (1*) *wowa_x_y_z__citation.txt*: citation file
      - additional files for data sets that have them
    - (1*) *wowa__full-corpus.xls*: all annotation data in one spreadsheet
    - (1*) *wowa__full-corpus.tsv*: all annotation data in one TSV file
  - (1*) *wowa__readme.pdf*: the file you are reading
  - (1*) *wowa__coding-guidelines.pdf*: coding guidelines used by contributors
  - (1*) *wowa__citation.txt*: citation file for the entire collection
  - (1*) *wowa__licence.txt*: licence file
  - (1*) *wowa__logo.txt*: WOWA logo
- (1*) **wowa__sources-pdf.zip**: source texts for all data sets
  - (41*) *x_y_z* folders, one for each data set
    - (N*) *wowa_x_y_z_#.pdf*: source texts, labelled *A*, *B*, *C*, etc.
- (23*) **wowa_x_y_z__sources-wav.zip**: source recordings for the data sets that have them
  - (N*) *wowa_x_y_z_#.wav*: source recordings, labelled *A*, *B*, *C*, etc.

Data sets

- *Armenian*

Armenian (Eastern, Agulis)	Katherine Hodgson

- *Hellenic*

Pontic Greek (Madan)	Katherine Hodgson
Pontic Greek (Romeyka)	Laurentia Schreiber

- *Indo-Aryan*

Kholosi (Kholos)	Maryam Nourzaei

- *Iranian*

Balochi (Coastal)	Maryam Nourzaei
Balochi (Koroshi)	Maryam Nourzaei
Balochi (Turkmen)	Geoffrey Haig
Bashkardi (Northern)	Agnes Korn, Ilya Gershevitch
Bashkardi (Southern)	Agnes Korn, Ilya Gershevitch
Gorani (Gawrajǔ)	Masoud Mohammadirad
Kumzari (Musandam)	Geoffrey Haig
Kurdish (Central, Sanandaj)	Masoud Mohammadirad
Kurdish (Northern, Ankara)	Kateryna Iefremenko
Kurdish (Northern, Lachin)	Donald Stilo
Kurdish (Northern, Muş)	Geoffrey Haig
Kurdish (Southern, Bijar)	Masoud Mohammadirad
Mazandarani (Kordxeyl)	Donald Stilo, Geoffrey Haig
Persian (New)	Elham Izadi
Persian (New, Early Classical)	Mehdi Parizadeh
Talyshi (Lerik)	Donald Stilo
Tāti (Hazārrudi)	Raheleh Izadifar
Vafsi (Gurchani)	Mahîr Can Doğan
Zazakî (Çewlîg)	Netîce Demir, Mahîr Doğan
Zazakî (Siwêreg)	Netîce Demir, Mahîr Doğan

- *Kartvelian*

Laz (Arhavi)	Donald Stilo, René Lacroix

- *Semitic*

Arabic (Jewish, Baghdad)	Assaf Bar-Moshe, Alexandru Craevschi
Arabic (Christian, Ka'biye)	Paul Noorlander
Arabic (Khuzestan)	Bettina Leitner
Central Neo-Aramaic (Mlahso)	Paul Noorlander
Central Neo-Aramaic (Turoyo, Midyat)	Paul Noorlander
NE Neo-Aramaic (Christian, Barwar)	Donald Stilo
NE Neo-Aramaic (Christian, Shaqlawa)	Paul Noorlander
NE Neo-Aramaic (Christian, Urmi)	Paul Noorlander
NE Neo-Aramaic (Jewish, Dohok)	Dorota Molin
NE Neo-Aramaic (Jewish, Sanandaj)	Paul Noorlander
NE Neo-Aramaic (Jewish, Urmi)	Paul Noorlander

- *Turkic*

Oghuz (Ankara)	Kateryna Iefremenko
Oghuz (Erzurum)	Mahîr Doğan
Oghuz (Gagauz)	Mahîr Doğan
Oghuz (Qashqai)	Sohrab Dolatkhah, Laurentia Schreiber
Oghuz (Tabriz)	Donald Stilo

References

Haig, Geoffrey & Noorlander, Paul & Schiborr, N. 2025. Which word order features are stable in a contact setting? Corpus-based evidence from the Western Asian Transition Zone. In Darquennes, Jeroen & Salmons, Joseph C. & Vandenbussche, Wim (eds.), *Language contact*: *An international handbook*, vol. 2, 159–184. Berlin: De Gruyter Mouton. (https://doi.org/10.1515/9783110443011).

Haig, Geoffrey & Stilo, Donald & Doğan, Mahîr C. & Schreiber, Laurentia & Schiborr, N. (eds.). 2024a. *Post-predicate elements in the Western Asian Transition Zone: A corpus-based approach to areal typology*. Berlin: Language Science Press. (https://doi.org/10.5281/zenodo.14185541).

Haig, Geoffrey & Stilo, Donald & Doğan, Mahîr C. & Schreiber, Laurentia & Schiborr, N. 2024b. Post-predicate elements in the Western Asian Transition Zone: Data, theory, and methods. In Haig, Geoffrey & Stilo, Donald & Doğan, Mahîr C. & Schreiber, Laurentia & Schiborr, N. (eds.), *Post-predicate elements in the Western Asian Transition Zone*: *A corpus-based approach to areal typology*, 3–54. Berlin: Language Science Press. (https://doi.org/10.5281/zenodo.14266331).

Noorlander, Paul & Molin, Dorota & Haig, Geoffrey. 2025. Between VO and OV in Arabic and Aramaic: A corpus-based typology with implications for word-order shifts. *Studies in Language* 49(2). 289–336. (https://doi.org/10.1075/sl.24027.noo).