

Tehlikedeki Diller Dergisi

Journal of Endangered Languages

e-ISSN 2148-130X

Documentation of Turkic in comparative linguistics

Lars Johanson

This article appears in: *Journal of Endangered Languages*, Year 2023, Volume 13, Issue 23. *Language Documentation in Comparative Turkic Linguistics*. Guest Editors: Éva Á. Csató & Birsel Karakoç

Cite this article:

Johanson, Lars (2023). Documentation of Turkic in comparative linguistics. In: Csató, Éva Á. & Karakoç, Birsel (eds.), *Journal of Endangered Languages 13, Vol. 23: Language Documentation in Comparative Turkic Linguistics*. Ankara. 1–6.

Link to this article:

<http://dergipark.org.tr/tr/pub/tdd/issue/82059>

This electronic version was first published in December 2023.



This article is published under a Creative Commons Attribution-NonCommercial 4.0 International License CC-BY-NC. The licence permits users to use, reproduce, disseminate or display the article, provided that the author is attributed as the original creator and that the reuse is restricted to non-commercial purposes, such as research or educational use.

See <http://creativecommons.org/licenses/by-nc/4.0/>

ETL Publishing articles and services

Websites: <https://dergipark.org.tr/en/pub/tdd>, <http://tehlikedekidiller.com/>

Policy of JofEL: <https://dergipark.org.tr/en/pub/tdd/aim-and-scope>

Author guidelines: <https://dergipark.org.tr/en/pub/tdd/writing-rules>

□ **Lars Johanson**

Johannes Gutenberg University Mainz, Germany

Documentation of Turkic in comparative linguistics

Abstract

This article argues for some reasonable and linguistically motivated standardization of transcription and terminology in the documentation of Turkic varieties. The suggested model is demonstrated by the case studies in this volume. The use of documented data in comparative Turkic studies requires transparency of typological features and identifiability of morphemic heritage. Selected examples illustrate the issues involved.

Key words: Turkic, language documentation, comparative approach, linguistic terminology, contact linguistics, code copying

Karşılaştırmalı Türk dilbiliminde belgeleme

Öz

Bu makale, Türk dillerinin belgelenmesinde çevriyazı ve terminolojinin makul bir şekilde ve dilbilimsel gerekçelerle standartlaştırılması konusunda görüşler ortaya koymaktadır. Önerilen model bu ciltte örnek çalışmalar ile sunulmaktadır. Karşılaştırmalı Türk dilbilim çalışmalarında belgelenen verilerin kullanılması, tipolojik özelliklerin saydam ve morfemik mirasın tanımlanabilir olmasını zorunlu kılar. Söz konusu sorunlar seçilen örnekler üzerinden açıklanmaktadır.

Anahtar sözcükler: Türk dilleri, dil belgelemesi, karşılaştırmalı yaklaşım, dilbilimsel terminoloji, temas dilbilim, kod kopyalama

Data on Turkic varieties

Turkic languages are commonly considered interesting because of their vast geographical distribution, their contacts with many different types of languages, their relative stability over time, and their regularity in morphology and syntax. However, through permanent differentiation, specialized kinds of spoken and written Turkic varieties have emerged. The family tree branches out into geographic varieties: dialect groups, regional dialects, and basic local dialects, and into social varieties, more or less important in terms of prestige. Applying a two-dimensional dialectology, we may delimit geographically and socially distinctive varieties that combine features belonging to particular areas and layers. They contrast horizontally with their neighbors, and vertically with other social varieties, including standard varieties. Each has a certain range of validity and communicative range.

Documentation of these varieties is a necessary condition for gaining more knowledge about common typological features and individual developments. However, parts of the Turkic linguistic world have so far been insufficiently investigated. Fieldwork is indispensable in many areas.

Fieldwork

Today we are facing new exciting possibilities for fieldwork, both politically and technically. The need for linguistic documentation is great. We not only need data from well-established Turkic languages, but also from less known vernacular varieties, peripheral languages, languages strongly influenced by contact, isolated languages displaying both archaic and innovative features, etc. Field research may result in further important first-hand data, which may again lead to considerable re-evaluations in Turkic linguistics. Linguistic documentation is an urgent task that is best carried out through international cooperation. It is exciting, particularly for speakers of the varieties, to take an active part in documenting languages and collecting new primary data. We need linguistically trained scholars and speakers who can produce empirically adequate and theoretically meaningful research. In this special issue of the *Journal of Endangered Languages*, the focus lies on comparative aspects of documentation, and the need for a reasonable standardization of transcriptions and terminology.

Language endangerment

Some of today's Turkic languages are endangered, or at least potentially endangered. The endangerment starts when younger generations begin to pay less attention to their ancestral code and switch over to a dominant code because they find it more useful, attractive, and prestigious. Currently, increasing endangerment and loss of languages is observed all over the world, a development that could lead to mass disappearance of languages and will extinguish variation in an irrevocable way. This makes documentation an urgent task. It is important to try to document endangered Turkic languages while it is still possible to do so. The Hans Rausing Endangered Languages Project at SOAS in London (2003–2015), under the leadership of Peter Austin, was dedicated to documentary linguistics, and played a role in addressing the problem of language endangerment, by providing training, archiving, publishing, and funding of projects. The work continues in Berlin at the Berlin-Brandenburgische Akademie der Wissenschaften, which provides grants for language documentation and houses the Endangered Languages Archive (ELAR). See also the journal *Language Documentation and Description*, the EL Publishing that publishes quality peer-reviewed open access eBooks and multimedia on documentation, description, and support of endangered and minoritized languages and the highly stimulating informative EL blog <http://www.el-blog.org/> (see also references in Austin this volume). These have also been useful to some of the Turkic communities whose languages are endangered and to the researchers who work with them. Similar support is needed also for future studies.

A special task of linguists is to train speakers of the lesser-known Turkic varieties to document language structures and use. In this volume, several articles are written by speakers who are engaged in maintaining their languages and cultural heritage. See, for instance, the article on the southern dialect of Kirghiz, or the article written by speakers of Bayat in cooperation with a linguist. Ideal is when the speaker is a well-trained linguist, as is the case in the Maku and the Finland Tatar articles.

Contact processes

Varieties have emerged as a result of internal development and complex contact processes with copying of foreign elements. Speakers of Turkic have taken copies into their own primary code and non-Turkic speakers shifting to Turkic have carried over copies from their primary code into the Turkic secondary code. Because of the unique mobility of Turkic-speaking groups, contact-driven developments have been especially important (Johanson 2000, 2021, 2023a).

Small Turkic languages in asymmetric, strongly dominated situations manifest high-copying tendencies. Note, however, that heavy code-copying does not cause code replacement and loss. Shift does not mean successive transition from one code to another through intermediary stages characterized by increasing copying. Speakers of a dominated code do not take over larger and larger parts of a dominant code, until they end up speaking it instead of the dominated one. In-

creasing influence on a dominated code does not lead to its abandonment. Languages are not abandoned for structural reasons. Consider the example of Karaim, which has been a full-fledged language for more than six hundred years in Lithuania, in spite of its strongly dominated status.

The decisive factors for shift are social and political in nature. Languages are abandoned because they are suppressed and no longer handed down. If social-political pressure leads to negative attitudes towards a dominated code, its speakers may decide not to transmit it to their children, who acquire it incompletely and, at best, grow up as partial speakers. Most members of the Karaim community shifted to the dominating Russian and Lithuanian languages, when the free exercise of religion was not possible and the language lost its primary function.

In the development of Turkic, we always have to reckon with contact situations in which mutually intelligible varieties have met and influenced each other. Such family-internal contacts were the normal situation in tribal confederations with their mobile heterogeneous groups. The encounters led to changes and the emergence of modified varieties (Johanson & Csató 2022b).

The proper analysis of contact processes requires a theoretical framework. The Code-Copying Model has proved to be an insightful and easily applicable model for describing contact influences in several Turkic and contact languages (Johanson 2000, 2023a).

Comparative approach to data presentation

In documentation and description of Turkic varieties it is recommended to take a comparative approach which highlights shared typological features and calls attention to deviating ones. See our recommendations about *Transcription and abbreviations* in this volume.

Throughout their history and in spite of their huge area of distribution, Turkic languages share substantial phonological, morphological, and syntactic characteristics. It seems justified to speak of a certain conservatism of the family, with a rather low rate of change. Structures have remained relatively stable over centuries. The characteristics of the Turkic linguistic type presented in several studies characterize even high-copying languages demonstrating many non-Turkic properties; see, for instance, Johanson (2021).

Some typical typological characteristics of Turkic phonological systems can be mentioned here in order to demonstrate the advantage of a comparative approach in documentation. The following examples serve to illustrate the use of a transcription system based on sound types. In this volume the data is presented in a Turcological transcription rendering them. This transcription is accompanied by audio files providing the evidence of phonetic details. In the Bayat article an IPA transcription is added.

The first example concerns vowel systems. Most Turkic languages exhibit eight short vowel types, *a, ı, o, u, e/ä, i, ö, ü*, which may be classified according to the three features front vs. back, unrounded vs. rounded, and high vs. low. The phonetic realization of these vowel types can be conspicuously different in different varieties. The typical systematic oppositions are, however, maintained. The Turcological transcription distinguishes between the vowel types *e* and *ä*. They differ with respect to degree of openness: *e* being less open than *ä*. These two types occur in most Turkic languages. The transcription does not imply that they are phonemes, and their phonemic status can vary. They are, for example, allophones in standard Turkish and Karaim, but separate phonemes in Azeri. The phonetic realization also varies in the varieties. The Turcological transcription captures the fact that there is a regular distinction between a less open and a more open *e* sound, which is a common characteristic of the Turkic type.

High vowels such as *i, ü, ı*, and *u* are pronounced lax in suffixes. Lax pronunciation is marked by a dot under the vowel sign and implies more opening, and tendencies for retractedness and centralization. Lax vowels occur also in stems, especially in Kipchak languages such as Kazakh. Their phonetic features are clearly different from those of tense vowels. However, their phonemic status as a rule corresponds to the status of tense vowels. For instance, a Karaim lax vowel in the verb stem *küi*- 'to laugh' represents a front high rounded vowel in the system and governs the quality of the

vowel in attached suffixes, as in the converb suffix in *kül-üp*. It is therefore neither insightful nor correct to transcribe the verb stem as *kul-*. Native speakers, when asked, do not identify this lax vowel with *u*. This example leads us to the next common Turkic typological feature, sound harmony.

The most general sound harmony phenomenon is intrasyllabic front vs. back harmony, which requires the segments of a syllable to be either front or back, e.g. Turkish *kül* <kül> ‘ashes’ vs. *kut* <kul> ‘slave’. This intersyllabic front vs. back harmony causes neutralization of the front vs. back distinction under the influence of a preceding syllable. If applied consistently, harmony rules yield exclusively front or back word forms, e.g. Turkish <ev-ler-im-e> <house-PL-POSS1SG-DAT> ‘to my houses’, <at-lar-ım-a> <horse-PL-POSS1SG-DAT> ‘to my horses’. Most languages also apply a rounded vs. unrounded harmony, which causes neutralization of the distinction rounded vs. unrounded in suffix high vowels, e.g. Turkish <el-im> <hand-POSS1SG> ‘my hand’, <gül-üm> <rose-POSS1SG> ‘my rose’. Certain languages also apply rounded vs. unrounded harmony to suffixes with non-high vowels, e.g. Kirghiz *üy-lör-üm-dö* <house-PL-POSS1SG-LOC> ‘in my house’. These common Turkic tendencies are transparent in the Turcological transcription. Phonetic details can disguise this principle that is undoubtedly relevant for understanding the system.

The contribution of consonants as signals of the frontness or backness of a syllable is also important, as, for example, the phonological opposition between front and back *k* sounds.

Whereas this opposition is a basic feature of most Turkic sound systems, the phonetic realization of the front *k* and the back *ķ* types is language-specific and varies in a great degree. In Turkish the front *k* and the back *ķ* types are distinguished in pronunciation but not in the script. The front variant is slightly palatalized and the back variant is velar, in dialects with a clearly more velarized pronunciation. In Karaim, the front variant is strongly palatalized, whereas the back variant is not palatalized. The strong palatalization of the front variant is the phonetic signal marking the opposition. In Azeri, the back variant is a fricative *χ*. Thus, the opposition is maintained in spite of different phonetic realizations. In the Turcological transcription a *k* stands for the front variant and *ķ* for the back variant. The use of sound types in descriptions, in addition to an account of the phonetic details supported by audio material, makes the common tendencies transparent.

An important Turkic typological feature is that nominal and verbal stems are sharply distinguished, with denominal and deverbal suffixes forming separate classes. There is a notational tradition in Turkic linguistics to mark nominal suffixes with + and verbal suffixes with -. Compare, for example, the Turkish denominal nominal marker {+}I} and the deverbal nominal marker {-(y)I}I}, both indicating occupation or profession, e.g. <süt-çü> ‘milkman’ ← <süt> ‘milk’, <sür-ücü> ‘driver’ ← <sür-> ‘to drive’.

Turkic languages possess rich morphological inventories with hundreds of bound derivational and inflectional markers. The verbal morphology comprises numerous categories expressing grammatical notions of actionality (*Aktionsart*), voice (passive, reflexive, causative, cooperative-reciprocal), deontic modality (possibility, impossibility, necessity), epistemic modality (e.g. presumption), evidentiality (indirectivity), negation, viewpoint aspect (intraterminal, postterminal, terminal), prospective, mood (indicative, imperative, voluntative, optative, hypothetical), tense (past), interrogation, and person-number. There is a wide variety of simple and compound aspect/mood/tense forms. Certain verbal markers have different functions in different syntactic positions. A case in point is Turkish {-mİŝ}, e.g. <Git-miŝ> ‘gone’ (nonfinite, verbal noun) versus ‘X has evidently gone’ (finite) (see Johanson & Csató & Karakoç & Nevskaya 2024).

Morphological glosses mark the identity of a morpheme, and consequently, preferably, should have the same cross-Turkic tag in spite of encoding functional differences. For comparative studies it is fruitful to make the historical identity of a morpheme transparent. The following examples will illustrate the issues implied.

The suffix called “Aorist” is present through the whole documented history of Turkic languages. The Turkish morphophonemic formula is {-(V)r} showing that this is the only suffix having both

low-vowel and high-vowel variants. The name “aorist” is cross-linguistically misleading, because it is used for a particular type of past tense in Indo-European languages. Nevertheless, it is a well-established term in Turkic linguistics and it would be unmotivated to replace it with a new term. In Turkic, it is an intraterminal form which has gained various functions in different Turkic varieties. Functional tags PRES (present tense), MOD (modality), PRO (prospective), FUT (future) occur in the literature. The form of the suffix has also undergone language-specific changes. The Chuvash marker is for instance $\{-(\emptyset)\}$ with dropping of the final stem vowel, e.g. *Vul-i-p* ‘I will read’. Functional differences and formal distinctions blur the identity of the suffix. In comparative Turkic linguistics the suffix should be tagged as AOR (aorist).

Finite verb forms are tagged as INTRA (intraterminal) or POST (postterminal), marking their viewpoint-aspect value (Johanson 2023b). The Turkish intraterminal in $\{-(\emptyset)/yor\}$ can function as a present tense form, but also in combination with forms marking anteriority. The identity of this suffix can be shown by assigning it the tag INTRA in its different usages, e.g. *Gel-iyor* ⟨come-INTRA⟩ ‘X comes’, *Gel-iyor-du* ⟨come-INTRA-DIST.COP⟩ ‘X was coming’, *Gel-iyor-muş* ⟨come-INTRA-EVID.COP⟩ ‘X has apparently come/ came’.

Another well-established Turkic term is “converb”. Ramstedt introduced it for a specific type of infinite verb form in Altaic languages. The cross-linguistic definition given by Haspelmath (1995: 3) “a nonfinite verb form whose main function is to mark adverbial subordination” does not cover the function of Turkic converbs, which can mark subordination without adverbially modifying the matrix predicate (see Johanson & Csató & Karakoç forthcoming). Thus, it is not felicitous to use a tag such as for example “verbal adverb” to replace “converb”. The tag CONV (converb) is recommended for comparative studies.

A basic typological Turkic feature is that pronouns in the third person are neutral for gender and animacy. A Turkish *o* can be translated as ‘he’, ‘she’ or ‘it’. This common Turkic feature is marked in translations as X, e.g. *Ok-uyor* ‘X reads’, emphasizing the Turkic grammaticalization of this pronoun.

Applications of the recommended standards

Several comprehensive publications have applied the standardizations recommended here.

The aim of using a Turcological standard is to provide data for comparative linguistic research in the Turkic language family and follow up in this respect the scholarly heritage of *Philologiae Turcicae Fundamenta* published in 1959. The editors of *Fundamenta*, Jean Deny, Kaare Grønbech, Helmuth Scheel, and Zeki Velidi Togan, created a remarkable volume on Turkic languages, a linguistic milestone in the study of Turkic. A great merit of *Fundamenta* is, as Togan put it, that it avoids a narrow-gauge Turcology that only deals with some selected languages. *Fundamenta* covers the whole Turkic language family. Its other crucial merit, thanks to Gerhard Doerfer’s editorial work, is that it is relatively homogeneous, with data from the individual languages presented in a comparable way.

A recent successor of *Fundamenta* is *Encyclopedia of Turkic Languages and Linguistics* (Brill) edited by the present author and an editorial team (Johanson 2023c). My long-standing research on Turkic languages has further developed comparative Turkic linguistics by drawing upon the insights of modern linguistic studies and providing my own innovative analyses, while demonstrating the importance of Turkic studies for linguistic typology. The *Encyclopedia* applies a common terminology for the description of the early stages of Turkic languages and their modern varieties. The guidelines, such as language descriptions, transcriptions, and notations, are based on works such as Johanson (2021, 2203a, 2203b, Johanson & Csató eds. 2022a). This guarantees a relatively high degree of homogeneity/comparability of the language descriptions.

The journal *Turkic Languages* (Harrassowitz) recommends the use of these same standards.

References

- Haspelmath, Martin (1995). The converb as a cross-linguistically valid category. In: Haspelmath, Martin & König, Ekkehard (eds.) *Converbs in cross-linguistic perspective*. Berlin & New York: Mouton de Gruyter. 1–55.
- Johanson, Lars (2000). *Structural factors in Turkic language contacts*. London: Curzon. Turkish translation by Nurettin Demir.
- Johanson, Lars (2021). *Turkic*. Cambridge: Cambridge University Press.
- Johanson, Lars (2023a). *Code Copying. The strength of languages in take-over and carry-over roles*. Leiden: Brill.
- Johanson, Lars (2023b). *Aspect in the languages of Europe*. Ankara: Nobel.
- Johanson, Lars (ed.) (2023c). *Encyclopedia of Turkic languages and linguistics*. Online. <https://brill.com/display/db/etlo?language=en>
- Johanson, Lars & Csató, Éva Á. (eds.) (2022a). *The Turkic languages*. London & New York: Routledge.
- Johanson, Lars & Csató, Éva Á. (2022b). Code copying and the strength of languages. In: Storch, Anne & Dixon, R. M. W. (eds.) *The art of language*. Leiden: Brill. 302–315.
- Johanson, Lars & Csató, Éva Á. & Karakoç, Birsal (forthcoming). Clause chaining in Turkic. In: Sarvasy, Hannah S. & Aikhenvald, Alexandra Y. (eds.) *Clause-chaining in the languages of the world*. Oxford: Oxford University Press.
- Johanson, Lars & Csató, Éva Á. & Karakoç, Birsal & Nevskaya, Irina (2024). 17. *The verb in Northeastern Turkic*. In: Vajda, Edward (ed.) *World of linguistics 10: The languages and linguistics of Northern Asia*. Berlin: Mouton de Gruyter. 919–1006.

About the author

Lars Johanson, one of the world's leading Turcologists, was born and educated at Uppsala University in Sweden. He is professor of Turcology at the University of Mainz, Germany, and senior lecturer at Uppsala University. He has published widely on descriptive and historical linguistics, mostly focusing on the Turkic language family. His book *Turkic* (2021) presents pioneering contributions to Turkic linguistics and language typology. He edits the journal *Turkic Languages* and the monograph series *Turcologica*. His recent publications include *Code Copying and the Strength of Languages in Take-over and Carry-over Roles* (Brill 2023), *Aspect in the Languages of Europe* (Nobel 2023). He is chief editor of *Encyclopedia of Turkic Languages and Linguistics Online* (Brill 2023).