Werk

Titel: Five dimensions of linguistic distance

Autor: Johanson, Lars

Ort: Wiesbaden

Jahr: 2018

PURL: https://resolver.sub.uni-goettingen.de/purl?666048797_0022|LOG_0011

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

Five dimensions of linguistic distance

Lars Johanson

Lars Johanson 2018. Five dimensions of linguistic distance. *Turkic Languages* 22, 36–42.

The article deals with dimensions of linguistic distance: genealogical distance, typological distance, lexicostatistical distance, intelligibility distance, and perceived distance. These dimensions should be kept distinct and studies on them should be dealt with as parallel lines of investigation. The relevant results can then be combined in order to obtain novel insights.
Keywords: linguistic distance, Turkic languages, lexicostatistics, dialectometrics

Lars Johanson, Institut für Slavistik, Turkologie und zirkumbaltische Studien. Johannes Gutenberg-Universität Mainz, Hegelstraße 59, DE-55122, Germany.
E-mail: johanson@uni-mainz.de

1. Genealogical distance

Genealogical distance can be demonstrated with the normal tools of comparative linguistics: regular sound-meaning correspondences.[1]

The Turkic languages and dialects form a well-defined family whose members are closely related to each other in the sense of genealogical proximity. We can clearly demonstrate that they are related, descended from a common ancestor.

The historical situation is strongly characterized by migrations. We are not dealing with settled populations with fixed geographical borderlines, as often in the linguistic history of Europe. The Turkic languages have historically diverged from each other, but many family members have also come to converge with each other through areal interaction. The dual forces of divergence and convergence have formed new clusters. Interaction in a number of contact areas has led to new constellations involving convergence and leveling. Some Turkic languages have served as koinés, regional or transregional lingua francas for intergroup communication, leveled varieties that also influenced other varieties within their respective areas of validity. Even the earliest kind of Turkic known to us might, in view of its transparent, regular structure, have been a koiné, a leveled language of this kind.

Speakers of Turkic varieties met each other in heterogeneous confederations consisting of various nomadic groups. The fronts changed continuously, though at irregular intervals. Related varieties did not occur in clear-cut geographic clusters.

1 The article is based on a talk at the panel *Five Dimensions of Distance in the Turkic Language Family* at the *Second European Convention on Turkic, Ottoman and Turkish Studies*, Hamburg, September 14–17, 2016. For references to relevant studies, see Csató & Menz in this issue.

Owing to demographic and political circumstances, floating nomadic unions of clans, tribes, and subtribes moved ceaselessly over huge distances. The intra-family contacts meant that groups using different codes were brought together to coexist in tribal confederations, in mixed speech communities based on new social networks. Abrupt reorganization processes led to the emergence of modified varieties. The varieties were sufficiently closely related to adapt to each other, that is, to undergo a certain leveling. Disparate varieties were knit together, came to resemble each other more closely, developed common features, and assimilated. An interesting modern example of this is the emergence of the Kashkay confederation in the province of Fars in Iran.

It is a popular misunderstanding that Turkic was a unified language until recently and that it was cut into pieces by language policy in the 20th century. Before that, linguistic continua existed in various parts of the Turkic world, with different languages and dialects intermingling without very well-defined boundaries. In the 20th century, a number of distinct standard languages were created with separate vocabularies, grammars, and orthographies. This raised new barriers that impeded written communication. It is also true that some varieties came under the umbrella of a standard language that did not correspond to their genealogical background.

One thing is still unclear: whether Turkic is part of a larger family, traditionally called Altaic. Turkic belongs to a distinct type represented by a transcontinental belt of areally adjacent Transeurasian (Turkic-Mongolic-Tungusic-Korean-Japanese) and Uralic (Finnic-Ugric and Samoyedic) languages, which share a number of basic structural traits, close similarities in phonology, bound morphology, and syntax. Most of these similar typological features do not provide any conclusive evidence for genealogical kinship, since they are known to be easily copied across languages. Some of the shared features may be attributable to general typological principles.

Turkic has been involved in a multitude of family-internal and family-external language contacts. Though language contact is sometimes thought to complicate the kinship picture, it does not invalidate the results of the genealogical classification. The Turkic languages exhibit specific linguistic core structures, which are not overruled by code-copying caused by their numerous external contacts. Extensive copying may make certain relations difficult to recognize, but it does not lead to dissolution of the family bonds.

In spite of all contact interactions, it is still possible to determine how closely related the family members are to each other. Thus, of the neighboring Turkic languages of the Volga region, Tatar-Bashkir-Chuvash, we can easily conclude that the first two are closely related to each other, while the third one is not. We can also demonstrate that the Oghuz languages Turkish and Turkmen are closer to each other than to Kipchak languages such as Kazakh.

2. Typological distance

Also typological distance, in principle independent of genealogical distance, may serve as the basis for classifying Turkic languages and varieties. Even closely related varieties may be relatively different from each other. Less closely related or unrelated varieties may become more similar to each other. For example, the Kipchak language Karaim and the Oghuz language Gagauz, both spoken on the western peripheries of the Turkic world, are typologically rather similar. Classifications based on both genealogical and typological distance may tell us much about the complex history of settlement of Turkic-speaking groups.

Throughout their history and across their huge area of distribution, Turkic languages show many shared core features. It seems justified to speak of a certain conservatism of the family, a relatively low rate of change. The oldest known Turkic variety, that of the East Old Turkic inscriptions of the 8th century CE, is in fact remarkably similar to modern Turkish. Its rich morphosyntax displays a high degree of regularity, maybe a result of early koinéization. Some salient typological characteristics occur across the whole family, with minor exceptions in languages such as Chuvash, Khalaj, and Yakut.

Some properties are commonly considered to be basic to Turkic structure. Most of them are found in the other Transeurasian languages.

With respect to relational typology, Turkic adheres to the nominative-accusative pattern. Its syntax is head-marking. It has a left-branching syntax with modifiers and dependents preceding their heads. The unmarked order of clause constituents is subject + object + predicate, with discourse-pragmatically and stylistically conditioned deviations. There are few instances of grammatical agreement. Omission of constituents such as subjects and objects is permitted if the referents are pragmatically recoverable, which includes pronoun-dropping ("null anaphora"). Main clauses mostly take on special finite markers. Non-main clauses are based on action nominals, participant nominals and converbs provided with non-finite bound junctors, largely fulfilling the functions of conjunctions in languages of the English type.

Case markers and postpositions of various kinds correspond to English prepositions. Grammatical gender is lacking. Nominal and verbal stems are sharply distinguished. Affixation is exclusively suffixal.

One characteristic of the agglutinative structure is a high degree of synthesis. The rich morphological inventories comprise hundreds of bound derivational and inflectional markers. A high degree of combinability allows bound markers to occur in long sequences. Another characteristic of agglutination is a juxtaposing technique with clear-cut morpheme boundaries. Bound morphemes mostly show phonologically predictable allomorphs. Turkic thus lacks phenomena such as different declinations or conjugations, irregular verbs, and suppletive forms.

The most general sound harmony phenomenon is the intrasyllabic front vs. back harmony, which requires the segments of a syllable to be either front or back. The intersyllabic front vs. back harmony causes neutralization of the front vs. back dis-

tinction under the influence of a preceding syllable. Most languages also apply a rounded vs. unrounded harmony, which causes neutralization of the distinction rounded vs. unrounded in high suffix vowels. Certain languages also apply this harmony to suffixes with non-high vowels.

There are numerous exceptions to the harmony rules, especially in languages under Iranian or Slavic influence. In fact, sound harmonies are only tendencies with unstable, volatile realizations. One erroneus assumption found in the Turcological literature is that the vowel harmony observed in today's Turkish is a constant, eternal property of Turkic. It is even thought that the oldest known Turkic language, East Old Turkic, had an identical harmony system. In reality, the present Turkish vowel harmony has developed since the 18th century.

Grammatical categories pertaining to the verb systems appear to be most fruitful for defining the specific core structures of Turkic: viewpoint aspect categories, post verbs as actional modifiers, moods, evidential markers. The verbal morphology comprises numerous categories expressing grammatical notions of viewpoint aspect (intraterminal, postterminal), actionality (Aktionsart), mood (indicative, imperative, voluntative, optative, hypothetical), evidentiality (indirectivity). There is a wide variety of simple and compound aspect/mood/tense forms. Even high-copying languages extremely affected by contact-induced changes maintain the rich aspect-mood-evidentiality menu. The typically Turkic categories have proven dominant in all contact situations. Chuvash, which relatively early left the bulk of Turkic languages and, on the surface, is very different from its relatives in the family, has preserved all the distinctions in the normal Turkic way. The so-called aorist, an old intraterminal category that has drifted into the modal domain, has been assumed to be lacking in Chuvash, which is definitely not the case.

Another interesting fact is, however, that Turkic languages typically renew their core categories by modifying their morphological expressions.

Many Turkic languages certainly exhibit exceptions to the typical features just mentioned. In particular, the clear-cut agglutinative structure is partly disarranged in Northeastern Turkic.

3. Lexicostatistical distance

Since the 1970s, researchers have endeavored to develop techniques for measuring the lexicostatistical distance between languages and dialects. This means that cognate words from basic vocabularies are compared with respect to the degrees of phonetic similarity. The purpose is to go beyond the traditional comparative methods in describing linguistic variation. The last decades have seen rapid developments in this field. Methods for calculating pronunciation distances between pairs of closely related language varieties based on words collected in different geographic areas have been highly successful in dialectometrical studies. One tool is the Levenshtein distance method, which measures the number of insertions, deletions, and substitutions that transform one phonetic string into another. Traditional investiga-

tions restricted to single features are supplemented by techniques to calculate aggregate distances.

Lexicostatistical distance is clearly distinct from genealogical distance, though wordlists of various kinds are often used to measure both. Counting look-alikes is not sufficient to prove kinship relations. Cognates do not necessarily look similar. Words in a pair of languages may be cognates without being recognizable as such. Reversely, without knowledge of the genealogical relations, languages such as Turkic Uzbek and Persian Tajik may be taken to be closely related.

Linguists have been carrying out interesting work on the preliminary identification and classification of languages in Central Asia, for instance in Ferghana, where the Turkic group consists of Kirghiz-Uzbek-Kazakh-Karakalpak, and the Iranian group of Tajik and Yaghnåbi. The task has been to calculate the distance between the pronunciation of a given word at geographically different sites.

The lexicostatistical methods may be useful in preliminary analyses of linguistic field notes when dealing with previously unknown areas of linguistic variation. In each case, it remains to be checked how well the aggregate distances match secured genealogical classifications.

The conditions for working with lexical items are not unproblematic. The lexicon is often liable to rapid and unpredictable changes, especially in situations of political and cultural transformation. Two languages may have been lexically very close to each other in the past, e.g. Turkish and Azeri. If one of them undergoes a language reform such as Turkish in the 20th century, the lexical distance between them suddenly increases.

The annotation of data is a problem, especially for less known languages and dialects. Measuring phonetic distances requires transcriptions. Relevant material is often not available in a form that readily lends itself for automatic analysis. Official orthographies are mostly too idiosyncratic to serve as a basis for comparisons. Very different orthographic forms may stand for similar phonetic structures. Thus, among the Turkic languages of Central Asia, Uzbek has an idiosynchratic orthography that is inspired by Tajik, and is highly incompatible with the spelling systems of Kazakh, Kirghiz, and Uyghur.

4. Intelligibility distance

Intelligibility distance is an interesting and much-debated issue that has opened a new sociolinguistic field of research. It investigates how well speakers of different languages and dialects, especially related ones, can understand each other without deliberately engaging in language studies. Research in this field should of course measure real understanding of utterances, both in situational contexts and outside of them. It should also, however, look at what people *say* they understand, what they *claim* to understand, and how they *act* upon utterances. Interlingual comprehension may be problematic even between languages that are genealogically and typo-

logically related, have a large common stock of cognates in their basic vocabulary, are lexixostatistically close to each other, and are geographical neighbors.

Phonetic and lexical distances obviously affect comprehension. Prosodic, morphological, and syntactic differences all play their roles. The chances to establish successful communication vary. Many problems of understanding depend on the specific topic. Typological proximity between the languages concerned may facilitate the communication. Even in cases of potentially high mutual intelligibility, unfamiliarity with habits of pronunciation may cause initial problems. As soon as listeners become accustomed to these habits, the level of comprehension may rise significantly.

Native speakers of different languages may, under certain conditions, practice what has been called "mother tongue talk in more than one language". They use their own "expressive" language for production and a "receptive" language for comprehension. Such modes of multilingual communication may lead to phenomena that were formerly called semi-communication. The comprehension is sometimes a one-way process, too asymmetric to be characterized as "mutual".

There are now important new efforts to investigate the possibilities and modes of multilingual communication, to describe the determinant linguistic and extralinguistic factors, and to identify cases of understanding, misunderstanding, partial understanding, guessing, and total incomprehension. A useful "pragmatic index of language distance" has been set up at the Middle East Technical University in Ankara.

It is far from clear how easily speakers of different Turkic languages understand each other. According to a widely accepted definition, forms of speech that are mutually intelligible are dialects of a single language. Turkic dialectology is however, still relatively weak and not very helpful. Mutual intelligibility between all Turkic-speakers of Turkic is of course out of the question, despite the never-ending claims that only one Turkic language exists in the world. On the other hand, continua are found across certain geographical areas, ranges of varieties that differ only slightly, often without precise borders, sometimes with transition zones in-between. In such chains, the differences may accumulate gradually such that speakers of the varieties A and B or B and C can understand each other, whereas speakers from the opposite ends of the chain, A and C, do not.

The role of genealogical closeness varies. Closely related and neighboring languages may be mutually intelligible, for example Bashkir and Tatar. There is considerable proximity between Kumyk and Karachay, Uyghur and Uzbek, Noghay and Kazakh, and so on. But speakers of geographically distant close relatives mostly have great comprehension problems. The closest relative of Yakut is Tuvan, but the ratio of intelligibility between the two languages is in fact near zero.

5. Perceived distance

Perceived distance between language varieties may be seen as a fifth dimension of linguistic distance, though it will not be dealt with here. This subjective distance is based on impressions at various linguistic levels, but not necessarily dependent on the degree of intelligibility. It can be measured in tests in which listeners judge on the degree of similarity between their own variety and other varieties.