Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2026”

June 24–26, 2026

Semantic profile of Turkish: macroshifts[i]

Gruntov Ilya
Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia
Hebrew University of Jerusalem, Jerusalem, Israel
altaica@yandex.ru

Abstract

This paper presents a systematic description of semantic shifts in Turkish based on comprehensive processing of a Turkish-Russian dictionary. To analyze the macrostructure of semantic variation, we introduce the concept of **macroshift,** a relationship between large semantic domains (taxa). Based on macroshifts, we construct a semantic **profile** of Turkish that shows the quantitative distribution of shifts between semantic domains. To our knowledge, this is the first attempt to create a complete semantic profile of an individual language based on systematic dictionary processing.

The work was conducted within the DatSemShift project — a database of semantic shifts in the world's languages. We describe the methodology for working with taxa and the principles of constructing weighted semantic profiles. The material consists of NA Baskakov's Turkish-Russian dictionary (48,000 words), from which 1,645 semantic shifts were extracted.

The proposed approach opens up prospects for comparative analysis of semantic profiles across different languages.

Keywords: semantic shifts, macroshift, semantic profile, polysemy, semantic domains, ontology, Turkish language, DatSemShift

DOI: 10.29003/2075-7182-2026-24-125-133

Семантический профиль турецкого языка: макропереходы

Грунтов И. А.
Институт языкознания РАН, Москва, Россия
Еврейский университет в Иерусалиме, Иерусалим, Израиль
altaica@yandex.ru

Аннотация

В статье представлен опыт систематического описания семантических переходов в турецком языке на основе полной обработки турецко-русского словаря. Для анализа макроструктуры семантического варьирования предложено понятие макроперехода — связи между крупными семантическими доменами (таксонами). На основе макропереходов построен семантический профиль турецкого языка, показывающий количественное распределение переходов между семантическими доменами. Насколько нам известно, это первый опыт создания полного семантического профиля отдельного языка на основе систематической обработки словаря.

Работа выполнена в рамках проекта DatSemShift — базы данных семантических переходов в языках мира. Описана методология работы с таксонами и принципы построения взвешенных семантических профилей. Материалом послужил турецко-русский словарь Н.А. Баскакова (48 000 слов), из которого извлечено 1 645 семантических переходов.

Предложенный подход открывает перспективы для сопоставительного анализа семантических профилей различных языков.

Ключевые слова: семантические переходы, макропереходы, семантический профиль, полисемия, семантические домены, турецкий язык, онтология, DatSemShift

1 INTRODUCTION

1.1 DatSemShift Database and Semantic Shifts

In the study of semantic changes in a diachronic and typological perspective, which is one of the important tasks of lexical semantics, in the last two decades there has been a significant shift from the analysis of individual cases and individual classes of polysemy to the systematic collection of cross-linguistic material. The emergence of large databases (DatSemShift [Zalizniak et al. 2012], CLICS [Rzymski et al. 2019], EvoSem [François et al. 2026] and others) opened up opportunities for studying typological patterns of semantic variation.

The present work is based on materials from the **DatSemShift** (https://datsemshift.ru [DatSemShift 2026]), the database which have been created and developed under the supervision of Anna A. Zalizniak since 2000. The database contains more than 10,400 semantic shifts, represented by more than 49,000 realizations from 2,331 languages of the world. Its important distinction from other databases on semantic typology is the fact that all shifts and realizations in the database were selected by experts, rather than obtained through automatic parsing.

A semantic shift is defined as a relationship of "cognitive proximity between two meanings A and B (or SOURCE-meaning and TARGET-meaning) that manifests itself in the fact that these two meanings are conflated within the limits of *one word in an extended sense"* [Zalizniak 2018: 772] or are included in a motivational model linking a word and its morphological derivative. **A realization** of a semantic shift is a specific word (or pair of words) from a particular language that illustrates this relationship. According to the approach adopted in the DatSemShift database, a semantic shift can also be illustrated by two cognate words from different languages, semantic evolution of the same word across various time periods, or borrowings from one language to another, provided that the semantic change that occurred in the process corresponds to this shift.

For example, the semantic shift 'moon' → 'month' is represented in the database by 457 realizations from 436 languages, including Adyghe (*maze*), Turkish (*ay*), Mingrelian (*tuta*), and many others.

1.2. From specific shifts to the macro level: problem statement

The accumulation of a large amount of empirical data (almost 5,000 unique semantic meanings in the database) allows us to move to a new level of generalization. Instead of analyzing individual specific shifts (' moon ' → ' month ', ' to hear ' → ' to understand ') it is possible to study the connections between semantic classes: how common are shifts from the ASTRONOMICAL OBJECTS domain to the TIME domain? From the PERCEPTION domain to the COGNITION domain?

For such an analysis, it is useful to classify shifts by **semantic domains** (taxa) to which the source and target meanings belong [1]. If two meanings A and B are in a semantic shift relationship, then the relationship between their semantic domains A ' and B ' is called **a macroshift.**

The complete set of macroshifts found in a language will be called the language's **semantic profile**. In a narrow sense, the semantic profile is the most stable part of the semantic profile, consisting of the most frequent macroshifts. The goal of this paper is to demonstrate what a complete semantic profile of a language based on macroshifts might look like and to demonstrate the methodology for constructing it using the example of the Turkish language.

2 Methods

2.1. Taxon system

To classify semantic meanings, the DatSemShift project team developed a hierarchical system of taxa — large-scale semantic domains. A total of 117 taxa are used, 62 of which are top-level taxa and 55 are lower-level taxa (subcategories).

[1] Strictly speaking, it only makes sense to talk about source and target meanings when we can determine the shift direction. This is possible in most cases, but not all. For example, out of 10,447 shifts in DatSemShift, the direction was determined for 8,796 (84.1%).

Full list of top-level taxa: Abstract Entities, Abstract Relations, Administrative Divisions, Age, Agriculture, Animals, Artifacts, Astronomical Objects, Body Parts, Cognition, Connectives, Discourse Markers, Diseases, Emotions, Ethnonyms, Evaluation, Existence, Festivals and Holidays, Food and Drinks, Geography, Human Beings, Human Qualities, Interjections, Internal Organs, Interpersonal Relations, Logical, Magic, Meteorology, Modal, Motion, Numerals, Parts, Perception, Phasal, Physical Actions, Physical Objects, Physical Processes, Physical Properties, Physiology, Plants, Politics, Populated Areas, Position, Possession, Pronouns, Proper nouns, Quantity, Religion, Scientific Terms, Semiotics, Sets of Objects, Social Actions and Relations, Social Institutions, Spatial Relations, Speech, Substances and Materials, Supernatural Beings, Text, Time, Transport, Units of Measurement, Volition.
Lower-level taxa are subcategories of higher-level taxa. For example :

- Animals. Mammals, Animals. Birds, Animals. Fish, Animals. Insects
- Physical Properties. Size, Physical Properties. Color, Physical Properties. Shape
- Human Beings. Kinship, Human Beings. Professions
  A complete list of all taxa can be found here: https://datsemshift.ru/compare_taxons

2.2. Principles of correlating meanings to taxa

Many concepts do not fit to a single semantic class, so each concept can be assigned from one to three taxa.

Examples :
latitiude: Geography | Scientific Terms.
pope: Human Beings | Human Beings. Social Status | Religion
adopted child: Human Beings | Human Beings. Kinship | Social Institution
illegal: Evaluation | Social Actions and Relations

Taxa are thus used as tags rather than as mutually exclusive categories. In the current version of the database:

- 2,860 meanings have only 1 taxon
- 1,847 meanings have 2 taxa
- 77 meanings have 3 taxa

2.3. Semi-automatic assignment of taxa

Manually assigning taxa to nearly 5,000 unique meanings is an extremely labor-consuming task. Therefore, the decision was made to use the ChatGPT API for automatically assigning taxa from a given list, followed by manual verification and error correction. The accuracy of automatic detection was about 88%.

2.4. Macroshifts and semantic profiles

Macroshifts

According to the definition given above, if two meanings A and B are in a semantic shift relationship, then the relationship between their semantic domains A ' and B ' is a macroshift. Moreover, if the meaning A has taxa { T1, T2 }, and the meaning B has taxa { T3, T4 }, then the semantic shift A → B corresponds to four macroshifts : T1 → T3, T1 → T4, T2 → T3, T2 → T4. However, if a meaning has both upper- and lower-level taxa (e.g., Animals and Animals. Mammals), then only upper-level taxa are taken into account when counting macroshifts. Lower-level taxa are indicated in brackets: Animals → Human Beings (Animals. Mammals → Human Beings. Kinship, Animals. Fish → Human Beings. Professions)

Standard and weighted profile

The above approach creates a "blurring effect": the more taxa a meaning has, the more macroshifts it generates. With the standard calculation, an implementation with a source meaning with n taxa and a target meaning with m taxa generates $n \times m$ macroshifts, which artificially inflates the profile. Therefore,

a weighted profile is proposed. In the proposed approach, each realization makes a single contribution to the profile, regardless of the number of taxa: the weight of a macroshift is calculated using the formula w = 1.0 / n × m, while subcategories of a single basic category (e.g., Animals. Mammals and Animals. Domestic Animals) are considered as a single basic category. This ensures that the overall sample size is preserved (the sum of the weights is equal to the number of realizations) and allows for more accurate comparisons between different datasets. Thus, each realization makes a total contribution of 1 to the profile, but this contribution is distributed evenly among all the generated macroshifts. A slight drawback of this approach is that, conversely, macroshifts generated by shifts in which the meanings have a minimal set of taxa receive a slightly higher weight.

In this paper, we present both the normal and weighted profiles of the Turkish language.

3 Material and Procedure

3.1. Selecting a source

To construct an example semantic profile, we selected **the Turkish-Russian dictionary edited by N.A. Baskakov** (1977), which contains 48,000 entries. This is a fairly large and detailed dictionary, providing good coverage of Turkish vocabulary across various strata (literary, colloquial, argot, etc.) and time periods. The Soviet school of lexicography ensured due attention to the various meanings of polysemous words.

The choice of a bilingual (Turkish-Russian) rather than a monolingual dictionary was motivated by practical considerations: bilingual dictionaries edited by the Soviet lexicographic tradition tend to provide systematic and granular coverage of polysemy, as each meaning requires an explicit Russian equivalent, which incentivizes lexicographers to distinguish meanings that might be conflated in a monolingual entry. At the same time, a theoretical concern arises: the semantic structure of the metalanguage (Russian) could in principle influence the delineation of meanings. A preliminary comparison with the main monolingual Turkish explanatory dictionary (TDK Sözlüğü) suggests, however, that the degree of divergence may be modest: for *çekmek* 'drag, pull' (which is the most polysemous verb in Baskakov's dictionary) both sources independently identify exactly 43 distinct meanings, suggesting that the meaning structure of highly polysemous items is relatively stable across bilingual and monolingual lexicographic traditions. A full systematic comparison of the two sources remains a desideratum, but this observation supports the reliability of the bilingual source as a basis for semantic profiling.

3.2. Criteria for selecting semantic shifts

All semantic shifts without exception that met the criteria of a “good non-trivial semantic shift” were extracted from the dictionary (see the list of criteria in [Gruntov, Rykov 2023: 163]).

The direction of each shift (source vs. target meaning) was determined primarily based on the ordering adopted in the dictionary: the first meaning listed was taken as the default source. However, it is acknowledged that dictionary order does not always reflect historical precedence. In ambiguous cases, additional criteria (etymology, historical data, degree of abstractness etc.) were applied. As noted in footnote 1, the direction was determinable for the vast majority of shifts in the database.

A total of 1645 semantic shifts were extracted.

3.3. Processing data

1. Extracting shifts from Baskakov's dictionary (manually)
2. Assigning taxa to all meanings (semi-automatically via ChatGPT API + manual verification)
3. Building up macroshifts according to the methodology described above
4. Frequency counting and visualization of the top 20 macroshifts

4 Results

4.1 General statistics

Turkish language (Baskakov's dictionary):

- Unique semantic shifts: 1 645 (including 973 polysemy cases, the rest are morphological derivation/compound words).
- Unique macroshifts : 626, with only 40 macroshifts represented by 10 or more specific shifts. Maximum number of shifts within a macroshift is 49 (Physical Actions | Social Actions and Relations).

Taxa for Source meaning	**Number of shifts**	**Taxa for Target meaning**	**Number of shifts**
Physical Actions	192	Human Qualities	117
Physical Properties	167	Social Actions and Relations	112
Artifacts	130	Evaluation	107
Body Parts	129	Human Beings	99
Animals	107	Animals	85
Human Beings	80	Artifacts	82
Motion	72	Physiology	68
Substances and Materials	66	Emotions	64
Physiology	55	Body Parts	60
Plants	41	Abstract Entities	60

Table 1. Distribution of taxa for shifts attested in Turkish (top 10)

It's interesting to compare this distribution with the distribution of taxa in the DatSemShift database as a whole (top 10 most frequent taxa by number of semantic shifts involved):

Taxa for Source meaning	**Number of shifts**	**Taxa for Target meaning**	**Number of shifts**
Physical Properties	952	Animals	855
Body Parts	918	Human Beings	660
Physical Actions	888	Artifacts	610
Animals	805	Social Actions and Relations	603
Artifacts	714	Plants	543
Human Beings	613	Body Parts	509
Substances and Materials	504	Human Qualities	482
Motion	503	Physiology	460
Plants	411	Evaluation	439
Physiology	369	Time	401

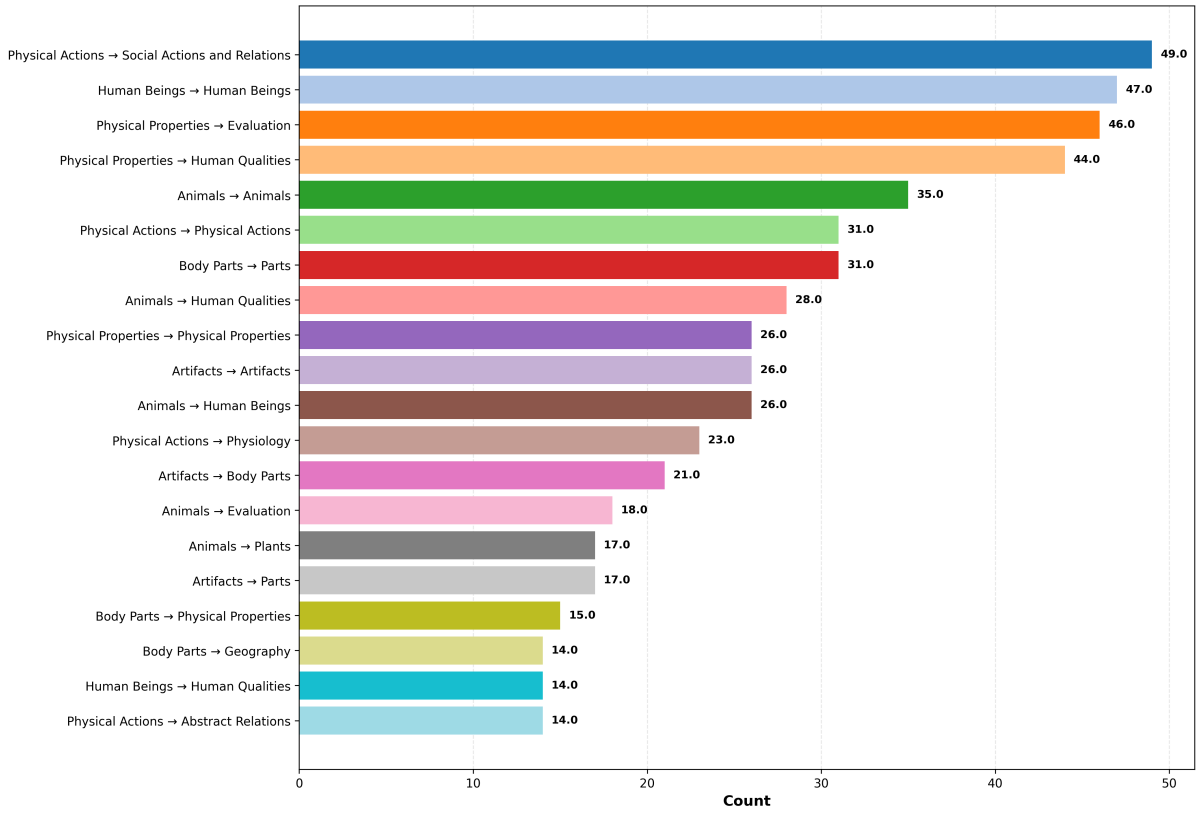
Table 2. Distribution of taxa in DatSemShift [2]database

A noticeable asymmetry can be seen: the taxa for the Turkish sources are generally very similar to the values in the database as a whole, while the taxa for the target differ drastically. A very high correlation is observed for the source taxa (Pearson's $r = 0.909$, $p < 0.001$; Spearman's $\rho = 0.891$, $p < 0.001$), indicating a virtually identical distribution of the conceptual domains that serve as the starting point for semantic shifts. All ten most frequent source taxa coincide in Turkish and in the database as a

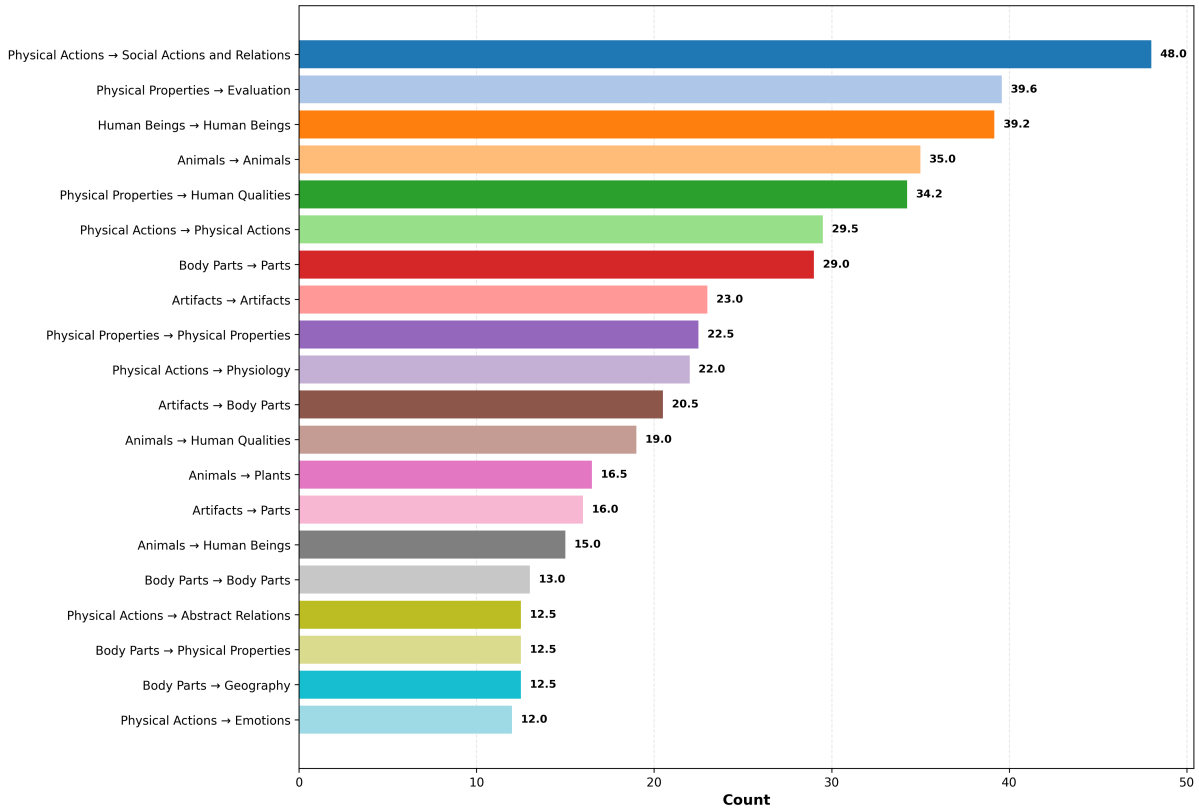
[2] Data taken from https://datsemshift.ru/stat?tab=overview , accessed February 13, 2026

whole, differing only by minor variations in proportions. On the contrary, for the target taxa there is no correlation ($r = -0.032$, $p = 0.94$; $\rho = -0.095$, $p = 0.82$): the Turkish profile is directed predominantly towards the abstract-social sphere (Human Qualities 13.7%, Evaluation 12.5%, Social Actions and Relations 13.1%, Emotions 7.5%, Abstract Entities 7.0%), while natural domains and temporal conceptualization dominate in the overall database (Animals 15.4%, Plants 9.8%, Time 7.2%). This asymmetry ($\Delta r = 0.941$) may reflect both the actual typological features of the Turkish language and be associated with a certain predominance of semantic shifts related to folk biology in the database. To test the hypothesis about the varying degrees of variability of sources and targets of semantic shifts, an analysis of other languages based on fully processed dictionaries is necessary.

4.2. Top 20 Macroshifts in Turkish (Standard Profile)



4.3. Top 20 Macroshifts in Turkish (Weighted Profile)



Comparison of the standard and weighted profiles shows that within the macroshifts of highest frequency, weighting affects mainly their relative order, but not their content.

4.4. Most productive macroshifts in Turkish

Intra-domain shifts :

- *Human Beings* → *Human Beings* (*çıtak* '1) stranger, outsideer, 2) braggart, liar '; *er* '1) man, 2) soldier, fighter', etc.)
- *Animals* → *Animals* (*göçken* '1) hare, 2) ermine'; *köpek* 'dog' -> *köpekbalığı* 'shark' (lit. dogfish), etc.).
- *Physical Actions* → *Physical Actions* (*açmak* '1) open, 2) turn on'; vurmak '1) beat, 2) kill'

These macroshifts reflect intra-domain polysemy: different types of animals, different types of people, variations of actions.

Cross - domain shifts :

- *Physical Actions* → *Social Actions and Relations*: yağlamak '1) to lubricate, 2) to flatter '; soymak '1) to undress, 2) to rob' etc.
- *Physical Properties* → *Evaluation*: *derin* '1) deep 2) serious '; doğru '1) straight, 2) correct ' etc.
- *Physical Properties* → *Human Qualities*:*hafif* '1) light (about weight), 2) frivolous, flighty ' etc.

The present study deliberately abstracts away from the question of the cognitive mechanism underlying each shift (metaphor, metonymy, semantic broadening, etc.) in order to focus on the domain-level architecture of the semantic profile. A qualitative examination of the top macroshifts suggests that the data are consistent with standard typological expectations. E.g., *Physical Properties* → *Evaluation* is predominantly metaphorical, while *Animals* → *Animals* involves primarily taxonomic extension and metonymic shifts, but a systematic classification of shift types within the macroshift framework is left for future work.

5. Discussion

To our knowledge, this study represents the first systematic identification of semantic macroshifts for a specific language based on a complete analysis of a lexicographic source. Macroshifts, as a holistic system within a semantic profile that characterizes the semantic organization of a specific language, have not previously been the subject of quantitative analysis.

Firstly, not all semantic macroshifts are equal for the language system. Macroshift *Physical Actions → Social Actions and Relations* (includes 49 shifts) is structurally more important for the Turkish than, for example, *Physical Actions → Meteorology* (includes 2 shifts). A quantitative picture allows us to identify this productivity hierarchy.

Secondly, intra-domain shifts play an important structural role. Traditionally, shifts between different conceptual domains (concrete → abstract, body parts → geographic objects, etc.) have been considered the most interesting. However, in the Turkish profile, shifts within a single taxon occupy the top positions: *Human Beings → Human Beings, Animals → Animals, Physical Actions → Physical Actions*. These shifts reflect the ways in which conceptual space is internally organized. For example, how kinship terms are extended to forms of address or professions, how age designations shift into kinship terms, and which animal names can shift into each other.

Third, clusters of macroshifts are revealed — systematic projections from one conceptual domain to several related target domains. Thus, *Physical Properties* become the source for a whole range of abstract and psychological concepts: *Evaluation* (39.58), *Human Qualities* (34.25), *Abstract Relations* (10), *Emotions* (10), *Cognition* (5.5) — a total of approximately 100 points of the weighted profile.

The main limitation of this study is the possible dependence of the profile on the choice of dictionary source. [Baskakov 1977], which served as the basis for creating the Turkish profile, is a fundamental academic dictionary, but any dictionary reflects not only the linguistic system but also the lexicographic priorities of its compilers. Reliable conclusions about the semantic profile of the Turkish language require a comparison of data from multiple dictionaries. If we can demonstrate that the profile is stable when using different dictionaries of the same language (explanatory, bilingual, complete, and concise), this will indicate that the semantic profile reflects the characteristics of the language as a whole, rather than the specifics of a particular lexicographic source. This, in turn, will open the possibility of creating a new, useful typological metric of linguistic diversity.

A second source of dependence is the ontological system itself: different taxon systems would yield different macroshifts, and it cannot be excluded that some findings are specific to the DatSemShift taxon inventory rather than the language. Testing the stability of semantic profiles under alternative ontologies (e.g., based on WordNet domains or CONCEPTICON) is an important avenue for validating the methodology.

To test the hypothesis that semantic profiles can serve as a reliable typological metric, it is also necessary to analyze profiles of other languages based on fully processed dictionaries and a unified methodology for filtering semantic shifts. Preliminary cross-linguistic comparisons based on the same methodology are available in the DatSemShift database (https://datsemshift.ru/compare_semantic_profiles), where semantic profiles of several languages can be compared visually. However, it is reasonable to compare languages with over 800 realizations of the semantic shifts in the database. These preliminary results suggest that languages do differ systematically in their target taxon distributions, supporting the hypothesis that the semantic profile reflects genuine typological characteristics rather than database artifacts. A full comparative analysis of profiles across multiple languages based on complete dictionary processing remains a priority for future work.

The main theoretical result of this work is the demonstration that semantic macroshifts form a system that can be quantitatively described and (subsequently) compared across languages. A semantic profile emerges as a characteristic of a language that reflects typical patterns of meaning change and relations between conceptual domains.

6 Conclusion

This paper proposes a method for constructing **a semantic profile of a language** based on the distribution of macroshifts between semantic domains. For the first time, a complete semantic profile of a single language (Turkish) has been created through systematic dictionary processing. Using a semantic

profile, it is possible for the first time to see the semantic architecture of a specific language at the macro level — that is, the system of pathways along which meanings flow from one conceptual domain to another. This architecture has a center and a periphery. It is quantitatively measurable and comparable across languages.

Main results:

1. the concepts of semantic profile and macroshift were introduced
2. A semi-automatic method for assigning taxa (accuracy ~90%) is proposed
3. The concept of a weighted profile has been introduced to correct the “blurring effect”
4. A semantic profile of the Turkish language has been built (1,645 shifts)
   The proposed methodology opens up prospects for comparative analysis of the semantic structures of the world's languages.

REFERENCES

[1] Baskakov A.N. (ed.) Turkish-Russian dictionary [Turecko-Russkiy slovar]. Moscow, Russkiy yazyk, 1977.

[2] DatSemShift 2026: Zalizniak Anna et al (2000-2026). "Database of Semantic Shifts". https://datsemshift.ru *Moscow: Institute of Linguistics, Russian Academy of Sciences.* [access date: 11-2-2026]

[3] Gruntov Ilya, Rykov Elisei, Computer-assisted detection of typologically relevant semantic shifts in world languages. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialog, [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: po materialam ezhegodnoy Mezhdunarodnoy Konferentsii “Dialog”], vol. 22, Moscow, RGGU, 2023, pp. 161-171.

[4] Rzymski, Christoph and Tresoldi, Tiago et al. 2019. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. DOI: 10.1038/s41597-019-0341-x

[5] François Alexandre, Dehouck Mathieu, Henke Konstantin & Kalyan Siva. (2026) 𝓔𝓿𝓸 Sem: A database of dialexification across language families. Online database. Lattice & HéLiCéO, CNRS - ENS–PSL, Paris. *https://tiny.cc/EvoSem* [access date: 11-2-2026]

[6] Zalizniak, Anna A. 2018. The Catalog of Semantic Shifts: 20 Years Later. *Russian Journal of Linguistics.*

[7] Zalizniak Anna A., M. Bulakh, D. Ganenkov, I. Gruntov, T. Maisak, M. Russo. The Catalogue of Semantic Shifts as a Database for Semantic Typology // M. Koptjevskaja -Tamm, M. Vanhove (eds.) New directions in lexical typology. Linguistics, a special issue. 2012. Volume 50, Issue 3, Pages 633–669.

[i] The work was supported by a grant from the Russian Science Foundation 22-18-00586-P (Mapping of the typology of polysemy via computer-assisted system of the cross-language identification of semantic shifts). https://rscf.ru/project/22-18-00586/