Russian Linguistics

The Structure of the Sight Domain: Slavic Verbs of Visual Perception in a Parallel Corpus

--Manuscript Draft--

Manuscript Number:	
Full Title:	The Structure of the Sight Domain: Slavic Verbs of Visual Perception in a Parallel Corpus
Article Type:	Original Paper
Funding Information:	Alexander von Humboldt-Stiftung (-) — Dr. Maria Ovsjannikova
Abstract:	This paper is a quantitative exploration of semantic relations between sight verbs in three Slavic languages – Bulgarian, Polish and Russian. Aiming at a maximally full coverage of this semantic class (113 verbs in total), I use the data of their pairwise correspondence frequencies in a parallel corpus. The goal is to identify the semantic distinctions structuring the domain and to explore the degree of semantic generality across sight verbs. Among the pairs of best-corresponding verbs, more frequent verbs have a higher degree of correspondence, suggesting that their range is not so much affected by semantic extensions and shifts as in the case of less frequent verbs. Sight verbs were found to be similar either to 'see'-verbs or to 'look'-verbs, or to neither of the two, testifying to the relevance of this distinction for the structuring of the sight domain. The similarity to basic sight verbs was also used to assess the degree of semantic generality of verbs. Another aspect of semantic generality was measured through the evenness of the distribution of correspondences of a sight verb of one language to verbs of another language. The results suggest that there is more potential for lexical renewal among 'look'-verbs as compared to 'see'-verbs. Finally, a tentative classification of semantic types of sight verbs was proposed, which includes, among others, verbs of temporally delimited perception, verbs of intense perception, and verbs of thorough perception.
Corresponding Author:	Maria Ovsjannikova, Ph.D. University of Potsdam: Universitat Potsdam GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Potsdam: Universitat Potsdam
Corresponding Author's Secondary Institution:	
First Author:	Maria Ovsjannikova, Ph.D.
First Author Secondary Information:	
Order of Authors:	Maria Ovsjannikova, Ph.D.
Order of Authors Secondary Information:	
Author Comments:	

Title

The Structure of the Sight Domain: Slavic Verbs of Visual Perception in a Parallel Corpus

Author information

Maria Ovsjannikova

University of Potsdam, Potsdam, Germany

maria.ovsjannikova@uni-potsdam.de

ORCID 0000-0002-8313-0374

Abstract

This paper is a quantitative exploration of semantic relations between sight verbs in three Slavic languages – Bulgarian, Polish and Russian. Aiming at a maximally full coverage of this semantic class (113 verbs in total), I use the data of their pairwise correspondence frequencies in a parallel corpus. The goal is to identify the semantic distinctions structuring the domain and to explore the degree of generality across sight verbs. Among the pairs of best-corresponding verbs, more frequent verbs have a higher degree of correspondence, suggesting that their range is no so much affected by semantic extensions and shifts as in the case of less frequent verbs. Sight verbs were found to be similar either to 'see'-verbs or to 'look'-verbs, or to neither of the two, testifying to the relevance of this distinction for the structuring of the sight domain. The similarity to basic sight verbs was also used to assess the degree of semantic generality of verbs. Another aspect of semantic generality was measured through the evenness of the distribution of correspondences of a sight verb of one language to verbs of another language. The results suggest that there is more potential for lexical renewal among 'look'-verbs as compared to 'see'-verbs. Finally, a tentative classification of semantic types of sight verbs was proposed, which includes, among others, verbs of temporally delimited perception, verbs of intense perception, and verbs of thorough perception.

Keywords

perception, lexical typology, parallel corpus, semantic generality

Acknowledgements

The study is part of the project "Perception verbs in Slavic: lexicon and grammar", funded by the Alexander-von-Humboldt Foundation. I thank the participants of the workshop "The Slavic verb – a multi-faceted approach", held at the University of Bamberg in April, 2024, as well as the participants of the Colloquium of Slavic linguistics department of the University of Potsdam for their comments and suggestions. The usual disclaimers apply.

Declarations

Funding

The reported study was funded by the Alexander-von-Humboldt Foundation.

Competing interests

The author has no conflicts of interest to declare that are relevant to the content of this article. The author has no financial or proprietary interests in any material discussed in this article.

Ethics approval/declarations

The research presented in the paper did not involve participation of humans or animals.

Consent to participate

Not applicable

Author Contribution

The paper as well as the study it was based on were carried out solely by the author.

Data availability

The files with the data used in this paper are available for download from: https://github.com/MariaOvsjannikova/Sight-verbs-in-Slavic

Materials availability

Not applicable

Code availability

The code used to process and analyse the data is available at https://github.com/MariaOvsjannikova/Sight-verbs-in-Slavic

The Structure of the Sight Domain: Slavic Verbs of Visual Perception in a Parallel Corpus

1. Introduction

Considerable advances have recently been made in the study of lexical semantics and colexification patterns of perception verbs (San Roque et al., 2018; Georgakopoulos et al., 2021; Norcliffe & Majid, 2024, among others). Due to their large cross-linguistic scale and the nature of the data, they are mostly restricted to basic perception verbs, such as *see* and *listen* in English. This study takes a close look at verbs of sight in three Slavic languages: Bulgarian, Polish, and Russian, which represent each of the three traditionally identified branches, i.e., Southern, Western, and Eastern Slavic, respectively. Using the data of the Intercorp Parallel Corpus (Rosen et al., 2022; Rosen, 2023), I analyse the frequencies of correspondences between a wide range of verbs of sight across the three languages to explore the semantic relations between them.

By taking a wider range of verbs into analysis, this study goes beyond the widespread paradigmatic view of the domain of perception (Viberg, 1984; Viberg, 2001). One of the structuring parameters of this paradigm is sense modality, which traditionally includes sight, hearing, smell, taste, and touch. For each of these sense modalities, the perceiver may either consciously direct their attention to an object, as exemplified by the English verbs *look* and *listen*, or perceive it without necessarily intending to do so, as in the case of the verbs *see* and *hear*. This distinction has been discussed under various terms, including activity vs. experience (Viberg, 2001), active vs. passive perception (Nesset et al., 2008), and opportunistic vs. explorative perception (Wälchli, 2016). In this paper, these two types of sight verbs will be referred to as ‘look’-verbs and ‘see’-verbs, to avoid the unnecessary associations with the domains of aspectuality and voice some of these terms suggest. Along with the verbs where the perceiver is in the subject position, for each sense modality there may also exist verbs with the perceived object in the subject position, such as English *sound*. The latter type of verbs will not be considered in the present study. The paradigm of perception verbs for Russian is discussed by Padučeva (2004: 204) and Divjak (2015).

Sight has been argued to be cross-linguistically the most prominent sense modality in terms of its textual frequency, lexical expression across languages, and the ability to develop non-perceptual meanings; see the hierarchy proposed by Viberg (1984; 2001), as well as Sweetser (1990: 39-40), San Roque et al. (2015; 2018). For the languages under study, as well as for other Slavic languages, the basic verbs of sight include two ‘look’-verbs and two ‘see’-verbs, with one imperfective and one perfective verb in each pair. Table 1 shows the basic verbs of sight in the languages under analysis. Example (1) shows a context where the basic imperfective verbs of sight correspond to each other in the three languages[1].

[1] All examples and, unless indicated otherwise, their English translations are taken from the InterCorp parallel corpus. The parallel contexts are given in the alphabetical order of languages (Bulgarian, Polish, Russian). The abbreviations BG, PL and RU stand for the respective languages throughout the paper.

Table 1. Basic verbs of sight in Bulgarian, Polish, and Russian

	‘look’-verbs		‘see’-verbs	
	Imperfective	Perfective	Imperfective	Perfective
Bulgarian	*gledam*	*pogledna*	*vidja*	*viždam*
Polish	*patrzeć*	*spojrzeć*	*widzieć*	*zobaczyć*
Russian	*smotret'*	*posmotret'*	*videt'*	*uvidet'*

(1) BG *Kăde e Kolja? – izvika toj, kato* ***gledaše*** *Kolja, bez da go* ***vižda****.*

PL *Gdzie jest Kola? – krzyknął,* ***patrząc*** *na Kolę i nie* ***widząc*** *go.*

RU *Gde Kolja? – vskričal on,* ***smotrja*** *na Kolju i ne* ***vidja*** *ego.*

‘Where is Kolja? he asked looking at Kolja and not seeing him.’ (author’s translation)

There are many sight verbs beyond the basic ones, but in the research on perception, the main focus has been on basic verbs. As a result, the diversity of perception verb types is largely overlooked and the semantic distinctions in the perception domain are represented as more discrete and uniform than they are in reality (Wälchli, 2016: 63-65). Wälchli (2016) looks at the distribution of perception verbs across parallel contexts to find the semantic groups of uses that arise from the data. Although different in terms of the specific methodology employed, the present study draws on his ideas of data-driven research based on parallel corpus data that looks at a wider range of perception types.

In parallel contexts, a verb in one language can correspond to a variety of verbs in other languages. For instance, in (2), there are three events expressed by sight verbs, and in each case, the basic verb of one language corresponds to non-basic verbs in the other two languages, e.g., the first event is rendered by BG *izgledam* ‘scrutinize’, PL *przyglądać się* ‘look closely’ and RU *smotret'* ‘look’, of which only the latter verb belongs to the basic sight verbs.

(2) BG *<…> otnačalo toj go* ***izgleda*** *ravnodušno, setne se nadigna i raztărka oči, ala kogato* ***pogledna*** *otnovo, veče ne go* ***vidja****.*

PL *<…> w pierwszej chwili* ***przyglądał się*** *z roztargnieniem, potem wyprostował się i przetarł oczy kułakiem, lecz kiedy* ***spojrzał*** *znów w to samo miejsce – nic tam już nie* ***zobaczył****.*

RU *<…> vnačale on* ***smotrel*** *nevnimatel'no, no potom sel i proter glaza, no kogda on* ***vzgljanul*** *snova, ničego ne bylo vidno.*

‘At first he stared at it listlessly, then he sat up and rubbed his eyes; but when he looked again he could not see it any more.’

The frequencies of correspondences between verbs in parallel contexts are taken to reflect similarity in their meanings and usage. In this study, these frequencies are used to establish the semantic similarity between sight verbs and their groupings in a bottom-up way. The goal of this study is to investigate the similarity relations between verbs of sight and identify the parameters relevant for the structuring of this domain. These similarities are explored at various levels of granularity, starting from pairwise correspondences between verbs to groupings within the whole range of sight verbs under analysis.

A large part of the study is concerned with the issue of semantic generality and the ways to quantitatively assess it. This notion has been defined differently depending on the domain of inquiry and is often left without an explicit definition. One of the possible definitions is that semantically general verbs “provide the central means by which humans are able to describe their experiences via a linguistic code” (Theakston et al., 2004: 62). In grammaticalization, semantic generality is typically understood as “abstract, schematic word meaning” (Hilpert & Correia Saavedra, 2017: 370), a characteristic of grammatical items as opposed to lexical. Under any of these views, semantic generality is expected to be associated with higher frequency. The studies on verb acquisition by Theakston et al. (2004) and on word dispersion in texts by Hilpert & Correia Saavedra (2017) aim to disentangle the effects of semantic generality and frequency. In the present study, frequency and semantic generality are also treated as separate properties of verbs, and two methods for assessing different aspects of semantic generality of sight verbs are proposed.

The paper is structured as follows. Section 2 gives an overview of the semantic distinctions and meaning extensions that play a role in the structuring of sight verbs or may affect the degree of their mutual correspondence. In section 3, I describe data retrieval and processing. Sections 4 through 7 present the results of the study, including the pairwise correspondences between the verbs under analysis, the similarity of various sight verbs to basic verbs of the class, the distribution of correspondences to non-basic verbs, and the semantic groupings of the verbs. Section 8 summarizes the main findings of the study.

2. Semantic distinctions and polysemy patterns of sight verbs

The goal of this section is twofold. First, it gives an overview of the distinctions that can be expected to play a role in the structuring of the sight domain, bearing in mind the groupings proposed in the literature and anticipating the results of the present study. Second, it discusses the polysemy patterns of sight verbs and shows how they can affect the correspondence frequencies in a parallel corpus.

One of the distinctions that play a central role in the study is that between basic and non-basic perception verbs. Basic verbs serve as the default means to describe a certain type of perception, and as such they are expected to have a higher token frequency as compared to other verbs denoting the same sense and to be semantically general, in the sense that they describe the respective perception type without specifying, e.g., its manner or duration (San Roque et al., 2015: 40). The basic Bulgarian, Polish and Russian sight verbs are given in Table 1 above. These verbs are indeed by far more frequent than the other sight verbs in these languages. Some non-basic sight verbs are as semantically general as basic verbs, in that they denote sight in a semantically general way, but they are more or less restricted in terms of register or grammatical features. One example is the Russian verb *gljadet'*, which is fully synonymous to the basic verb *smotret'* ‘look’ and has a very similar distribution of grammatical forms, except that *gljadet'* is much more frequently used as a converb (*gljadja*) and is mostly attested in fiction. Other non-basic verbs can also be more or less similar to basic verbs, which manifests in the frequency of their correspondences to basic verbs in other languages. Similarity to basic verbs is discussed in section 5.

As discussed in the introduction, basic sight verbs are further subdivided into ‘look’-verbs and ‘see’-verbs. In works on perception, this distinction is often taken to be based on the presence vs. absence of control but in reality, it may be more complex (Wälchli, 2016: 71-72). In particular, the variation observed in the distribution of the two types of verbs in parallel contexts suggests that this distinction is not realized uniformly across languages and that there are types of perception events prone to more variation, such as ambulatory vision (‘go/come and see’) discussed by Wälchli (2016: 72, 79). The present study reconsiders this distinction throwing a wide range of non-basic sight verbs. In anticipation of the results presented in section 5, non-basic verbs can be classified as similar to either ‘look’-verbs or ‘see’-verbs, or neither of the two. However, there are no verbs which are intermediate between the two types. This study thus confirms the relevance of this distinction for the overall structuring of the sight domain.

The fact that basic verbs denote sight in the most general way implies that they can be used in place of a wide range of more semantically specific verbs, e.g., RU *smotret'* can replace such verbs as *ustavit'sja* ‘stare’, *ogljdyvat'sja* ‘look around’ or *razgljadyvat'* ‘examine’. In parallel texts, this manifests in the ability of a verb with a more general meaning to correspond to a wider variety of verbs in another language. This makes the distribution of their correspondences to the verbs of another language more even. Evenness of the distribution, discussed in section 6, is regarded as another aspect of semantic generality, along with similarity to basic verbs.

Many of the non-basic sight verbs in Slavic are morphologically prefixed verbs, which belong to various Aktionsarten (Padučeva, 2004: 198; Nesset, 2010). The semantic components that these verbs bear include brief and long perception, as well as intensive and complete perception. Wälchli (2016: 56, 99) discusses obscured perception verbs, such as RU *razgljadet'*, which denote seeing something despite bad conditions. Prefixed verbs also typically express various spatial configurations accompanying sight, such as looking around, over, or from behind an obstacle. The same prefixed verb can often highlight different aspects of the perception event at the same time. For example, BG *ogleda* can convey both the meaning of looking around (3) and fully perceiving an object (4), and often it is not easy to distinguish between the two.

(3) BG *Toj nervno* ***ogleda*** *pustite xălmove.*

PL ***Rozejrzał się*** *nerwowo po okolicznych pustych wzgórzach.*

RU *Leonardo ispuganno* ***ogljadelsja*** *po storonam.*
‘He looked nervously around the deserted hills.’

(4) BG *No posle* ***ogledax*** *lodkata im.*
PL *Ale* ***przyjrzałem się*** *ich łodzi.*
RU *Potom ja xorošen'ko* ***prismotrelsja*** *k ix lodke.*
‘Then I got a good look at their boat.’

Due to the high number of prefixed verbs and the differences in their usage patterns across the three languages, establishing a clear typology of these verbs is challenging. For this reason, investigating the groupings of these verbs that emerge from the data can be particularly important and insightful, as shown in section 7.

Among the more general semantic and grammatical distinctions structuring the domain of sight, aspect is the most prominent. Differences in aspectual behaviour can naturally affect the distribution of verbs across contexts, and thus the degree of correspondence between verbs across languages (Wälchli, 2016). Generally, this study does not delve into the differences in the aspectual behaviour of individual verbs. However considerable they may be, the study shows that, with very few exceptions, aspect remains one of the major parameters in categorizing the verbs under analysis. Specifically, a higher degree of correspondence is consistently observed between verbs of the same aspect. This may reflect the general patterns of aspectual usage in these languages which are not restricted just to sight verbs. At the same time, this generalization may not hold as strongly for Slavic as a whole, because the selected languages do not lie on the opposite poles in terms of aspect usage. In particular, in the east-west split suggested by Dickey (2000), both Bulgarian and Russian belong to the east group and Polish is considered an intermediate between the two; see also von Waldenfels (2012) on the use of aspect in imperative.

Polysemy patterns and meaning extensions are usually discussed in relation to basic verbs of perception (San Roque et al., 2018; Georgakopoulos et al., 2021). They are relevant for the present study, because the meanings expressed by basic verbs of sight in one language can be taken up by non-basic verbs in another language. For instance, one of the polysemy patterns cross-linguistically attested for perception verbs is when these verbs are employed to express the meaning of co-identification, which can be paraphrased by English expressions *consider to be* or *regard as* (San Roque et al., 2018: 380). The parallel corpus data suggest that in Bulgarian and in Polish, the basic ‘look’-verb is more widely used is this meaning, whereas in Russian, it is mostly expressed by the non-basic verb *rassmatrivat'* ‘examine’, as in (5).

(5) BG ***Gledame*** *na ubijstvoto na Marks kato na spănka.*

PL ***Patrzyliśmy*** *na śmierć Marksa jak na komplikację.*

RU *My* **rassmatrivaem** *smert' Marksa kak prepjatstvie.*

‘We've been looking at Marks ' death like it 's a setback.’

Meaning extensions of sight verbs to other domains, such as cognition (e.g., ‘understand’, ‘deduce’) and attention (e.g., ‘examine’, ‘check’) (San Roque et al., 2018: 380), can manifest in correspondences between these verbs and verbs of other semantic classes in parallel contexts. For some verbs, more than half of the occurrences in parallel texts fall on the correspondences with verbs outside of the domain of sight, e.g., RU *rassmatrivat'* often corresponds to Bulgarian and Polish verbs with the meanings ‘examine’, ‘consider’ and the like. In section 3, I show to what extent the occurrences of the sight verbs in each of the languages overlap with the sight verbs in other languages.

Setting aside the recognized extensions of meaning, situations involving sight proper in reality encompass a variety of subtypes, such as watching a movie, looking into a mirror, looking into someone’s eyes, or more abstract contexts like looking into someone’s soul or heart or looking into the future. In a particular language, some of these minor subtypes of sight can call for a specific verb, distinct from the basic one. This also affects the degree of correspondence between verbs in parallel texts. Consider a situation of looking out of the window as an example, see (8).

(8) BG *Tja se ražoždaše iz kabinata, nadničaše văv vsički ăgli,* **pogledna** *prez prozoretsa*

PL *Čodziła po kabinie, zaglądała we wszystkie kąty,* ***wyjrzała*** *przez okno*

RU *Ona hodila po kabine, zagljadyvala vo vse ugly,* ***posmotrela*** *v okno*

‘She walked around the cabin, looked into all corners, looked out of the window.’ (author’s translation)

To explore the correspondences between sight verbs for this situation type, I used a trilingual subcorpus of the InterCorp to search for the contexts containing BG *prez prozoretsa*, PL *przez okno*, and RU *v okno*. Then, I manually selected the contexts describing the situation of looking out of the window from inside and annotated them for the sight verbs used in each of the languages. The major correspondences between verbs attested in this dataset are shown in Table 2. The upper row shows Polish verbs, and the columns show the frequencies of their correspondence to Russian and Bulgarian verbs (the correspondences between the latter two languages are not shown). The verbs are ordered by frequency; the cells for perfective verbs are in grey.

Table 2. Verb correspondences for the situation of looking out of the window

		PL	

		wyjrzeć	*patrzeć*[2]	*wyglądać*	*spojrzeć*	*gapić się*	*popatrzyć*	*zaglądać*	**Sum**
BG	*pogledna*	39	2		6		2		**49**
	gledam		15	8	1	3		2	**29**
	zagledam se	1	1	1	1		1		**5**
	pogleždam	2		2					**4**
	zjapam		1			1			**2**
	nadničam			1	1				**2**
	vziram se					1			**1**
	Sum	**42**	**26**	**12**	**9**	**5**	**3**	**2**	**99**
RU	*vygljanut'*	26	2		5		1		**34**
	smotret'	3	19	8		1		1	**32**
	posmotret'	12	3	1	4			1	**21**
	gljadet'		1			2	2		**5**
	ustavit'sja		1	2					**3**
	pjalit'sja					2			**2**
	vygljadyvat'	1		1					**2**

The majority of uses in Table 2 are distributed between basic ‘look’-verbs and verbs which convey an idea of looking from behind an obstacle. The three languages behave differently with respect to the distribution between these types. Polish prefers verbs specifying spatial configuration, especially for the perfective aspect (*wyjrzeć*). In Bulgarian, semantically general verbs are predominantly used in this context, whereas the other verbs are much less frequent. Finally, the distribution in Russian is more similar to that in Polish: among perfective verbs, the semantically specific verb *vygljanut'* is preferred, but the imperfective *vygljadyvat'* is less frequent than its Polish counterpart *wyglądać* (as compared to the respective perfectives; two-tailed Fisher test, p ≈ 0.04).

Table 2 also shows that the only feature where the correspondences between verbs are relatively consistent is aspect: imperfectives and perfectives typically correspond to verbs of the respective aspect in another language. Otherwise even within this semantically restricted situation type there is considerable variation in the choice of verbs.

This example served to show how correspondences between individual sight verbs result from the patterns of correspondence in particular situation types. It is thus unsurprising that when the whole range of uses is considered together for several dozens of verbs in each of the languages, the emerging picture is even more complex and multifaceted. It also suggests that in the light of parallel corpus data, no categorical distinctions, be it between different types of verbs or between different meanings of the same verb, can be expected to be found. However, one can expect to find frequency patterns indicating

[2] The counts for the Polish verb *patrzeć* here and elsewhere in the study also include the search results for the lemma *patrzyć*, which is regarded as its orthographic variant.

that some distinctions are more clear-cut and consistent that others. It is the aim of the present study to explore and identify these distinctions.

3. Data retrieval and processing

The major source of data for this study is the parallel corpus InterCorp (Rosen et al., 2022). The searches were conducted online using bilingual parallel subcorpora, i.e., each time searching in all the texts available only in two of the three languages. Although subcorpora of different sizes and slightly different sets of text types are available for the three pairs of languages, most texts in all the three subcorpora are subtitles, as shown in Table 3.

Table 3. Size and composition of the three bilingual subcorpora in InterCorp (as of September, 2024)

Language pair	**BG-PL**	**BG-RU**	**PL-RU**
Size	223.5 mln tokens	115 mln tokens	136 mln tokens
Composition	Subtitles 85% Legal texts 8% Discussions’ transcripts 5% Fiction 3%	Subtitles 96% Fiction 4%	Subtitles 92% Fiction 7% Other 1%

As pointed out by Levshina (2016: 516), subtitles represent spoken discourse and spontaneous conversations, which are only marginally present in fiction and other text types. For the present study, this means that the data may contain more discourse uses of sight verbs, e.g., imperative forms employed to direct attention or manage interaction (San Roque et al., 2018). In other respects, parallel subtitles have been shown to be a reliable source of data for language comparison, despite their translational nature and the specific conditions of their creation and use (Levshina, 2017).

Data retrieval and processing included the compilation of the lists of verbs, the retrieval of frequencies for all the pairs of verbs, and the correction of the data for false correspondences.

First, a list of verbs was compiled for each of the languages under analysis. The lists include experiencer-subject verbs describing visual perception.[3] The lists were mostly compiled bottom-up, based on the verbs attested in the searches. The lists were intended to be as exhaustive as possible. However, even apart from the fact that it is impossible to create a truly exhaustive list of any semantic class, as new verbs can emerge in language use without being at once or ever covered by any corpus, a frequency threshold had to be introduced to ensure statistical reliability of results. Only verbs having more than 100 occurrences in all the bilingual corpora were included. As a result, the lists included 34

[3] The lists do not include the verbs denoting lack of perception, such as PL *przeoczyć* ‘overlook’ and the verbs with the nouns denoting eyes as the direct object, such as BG *vtrenča* ‘stare’.

Bulgarian, 37 Polish, and 42 Russian verbs.[4] The verbs are given in the supplementary material 1, together with their frequencies in the respective monolingual subcorpora of InterCorp.

At the next stage, the frequencies of correspondence between all pairs of verbs for the three pairs of languages were found, using lemma search. Unfortunately, the frequencies yielded by the searches could not be taken at face value, and additional processing was necessary to determine the number of correct correspondences. The general criterion for identifying a pair of parallel contexts as a correct correspondence was the identity of construction type, i.e., experiencer-subject, and the semantic identity of participants. The latter is especially relevant for cases where the perceived object is expressed in different ways in the two languages, as in (3)-(4) above[5].

Some hits shown by the corpus were false correspondences. Most frequently, these are sentences featuring two sight verbs used in a sequence of perception events, where one verb describes direction of attention and the other the resulting perception event, as in (6). Another frequent configuration leading to false correspondences features one experiencer perceiving the perception event by another experiencer, as in (7).

(6) BG *Frodo* ***se ogleda*** *nazad i* ***zărna*** *otbljasăka na bjala pjana sred sivite dărvesni stăbla.*
PL *Frodo* ***obejrzał się*** *i* ***dostrzegł*** *blask białej piany między szarymi pniami drzew.*
RU *Frodo* ***ogljanulsja i uvidel*** *sredi drevesnyx stvolov beluju penu vodopada.*
'Frodo looked back and caught a gleam of white foam among the grey tree-stems.'

(7) BG *I dokato nikoj ne* ***gledaše***, ***vidjax*** *kak tja* ***pogledna*** *kăm teb.*
PL *Gdy nikt nie* ***patrzył***, ***widziałem***, *jak na ciebie* ***zerka***.
RU *I kogda nikto ne* ***videl***, *ja* ***zametil***, *kak ona* ***smotrit*** *na tebja.*
'And when no one else was looking, I saw the way she glanced at you.'

To correct the data for false correspondences, the following procedure was implemented. If the number of hits for a pair of verbs was less than 100, the examples were manually analysed and the exact number of correct matches was counted for the pair. If the number of hits was greater than 100, the search results were shuffled and the proportion of correct hits was counted for the first 100 occurrences. Then the number of hits shown in the corpus was multiplied by that proportion.

Additional processing of corpus results is also necessary in case of verb stems which can be used both with and without a reflexive marker, such as BG *zagleždam* (*se*). In this respect, one of the

[4] This difference in the number of verbs for the three languages might be due to the differences in the number of stems employed in the sight domain and the degree to which they are used to derive doublets. For instance, in Russian, the stems *smotret'* and *gljadet'* 'look' serve as a basis for many almost synonymous prefixed and reflexive verbs.

[5] Thus, valency patterns as far as non-subject participants were concerned were not taken into account when identifying correspondences between the verbs.

grammatical differences between Bulgarian and Polish, on the one hand, and Russian, on the other, is that in the latter two languages, the reflexive marker is a clitic written separately from the verb, and in Russian this marker is a suffix which is written together with the verb. Therefore, for Russian, the same stem with and without the reflexive marker is treated as two different lemmas in the corpus. For the other two languages, the implemented approach was to estimate the number of reflexive and non-reflexive uses based on the first 100 examples or count their frequencies among all the examples, if less than 100.

Reflexive markers in all the three languages have a wide range of functions, including reflexive proper (in a broader or narrower sense), reciprocal and passive (Geniušienė, 1987; Knjazev, 2007). Reciprocal and passive verbs were not included in the list for Russian and such uses were excluded from the counts of correspondences for Bulgarian and Polish. Note that this did not affect passive participles, as participles are retrieved in the searches of the verbs as their morphological forms.

The final dataset includes three tables, each containing raw frequencies for all pairwise correspondences between verbs for one of the three pairs of languages. Table 4 shows an excerpt of the table for Bulgarian and Russian verbs, given in columns and rows, respectively. The frequencies with decimal points resulted from correcting for false correspondences for verbs with more than 100 hits.

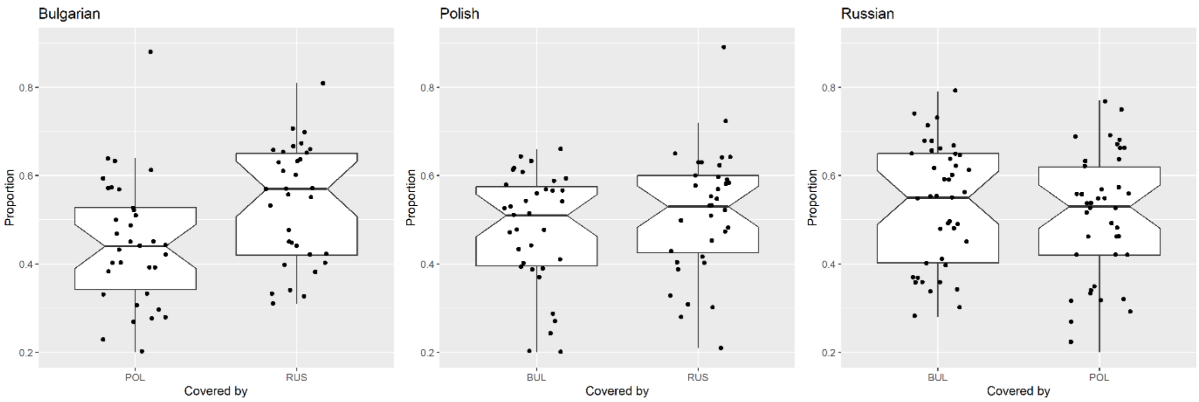
Table 4. Excerpt of the table with correspondences between Bulgarian and Russian verbs

Verbs	*videt*	*smotret*	*posmotret*	*uvidet*	*zametit*	*vzgljanut*	*nabljudat*
vidja	48667.78	9932.32	23212	30612.23	2137.59	3326.12	287
viždam	49482.24	1463.2	262.96	1975.24	338.56	30.75	237.9
gledam	3051.18	19646	3781.44	742	34.02	204	1592.5
pogledna	312.3	3975	11381.72	258.23	9.6	3748	16
zabeleža	657.93	30.88	94	450.34	5987	10	17
nabljudavam	191.7	867.24	131.58	78.5	81	12	2086.42
pregledam	31	44	321.6	11	1	128.7	3

For each verb, the sum of frequencies of correspondence to the verbs in another language is lower than the overall frequency of this verb in a given bilingual subcorpus. For instance, for RU *videt′* ‘see’, the sum of the frequencies of correspondence to all the Bulgarian verbs included in the dataset is 102890.06, whereas its overall raw frequency in the Bulgarian-Russian bilingual subcorpus is 159141. Thus, the proportion of the uses of *videt′* “covered” by Bulgarian verbs is 0.65. In the remaining 0.35 of its uses, *videt′* might correspond either to other Bulgarian sight verbs, not included in the list, or to verbs from other semantic domains. Recall that a different number of verbs is included for the three languages. Comparing these proportions for the three pairs of languages can help assess whether this difference affects the degree of “coverage” of verbs of one language by the verbs of the other in each pair. These proportions were calculated for each verb in the three pairs and visualized using boxplots, as shown in

Figure 1[6]. For instance, the leftmost plot shows the proportions of uses of Bulgarian verbs covered by Polish (on the left) and Russian (on the right) verbs. The line in the middle of each box shows the median of proportions; the box itself indicates the interquartile range, with one quarter of the calculated proportions lying below, and one quarter above the median (Kabacoff, 2011: 133-137). The notches correspond to the confidence interval around the median, and when they do not overlap, this indicates that there is a statistically significant difference between the two distributions. The points correspond to individual verbs.

Figure 1. Proportions of uses of the verbs covered by the sight verbs of the other language for the three pairs of languages

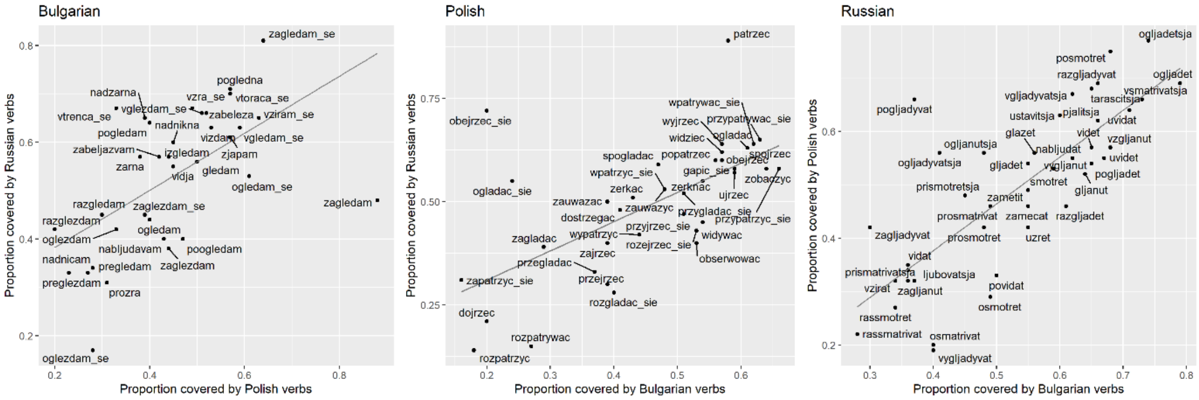


As shown in Figure 1, a statistically significant difference is observed only for the proportions of coverage of Bulgarian verbs by Polish and Russian verbs. This means that, as compared to the Polish verbs, the Russian verbs cover the Bulgarian verbs better. One of the reasons could be that there are simply more Russian sight verbs than Polish ones. However, in the other two cases, the medians of the distributions are very close, which suggests that the number of verbs as such is not significant. It is especially noteworthy that, as shown by the rightmost plot, the 34 Bulgarian verbs cover the 42 Russian verbs slightly better than the 37 Polish verbs. This suggests that Russian and Bulgarian may be more similar to each other in terms of the range of uses of sight verbs than to Polish. More importantly for this study, there is no consistent evidence that the difference in the number of verbs included in the lists for the three languages affects the proportion of uses covered by them in the other languages.

For each language, there is a strong correlation between proportions of the verbs' uses covered by the verbs of the other two languages (r(32) = 0.57, $p < 0.0004$ for Bulgarian, r(35) = 0.64, $p < 0.0001$ for Polish, r(40) = 0.78, $p < 0.0001$ for Russian). This means that, for instance, individual Bulgarian verbs are covered by Polish and Russian verbs to a similar extent. The correlations are shown in Figure 2.

[6] All the calculations and visualizations in this paper were performed using R (R Core Team, 2023). Unless indicated otherwise, the visualizations were created using the package ggplot2 (Wickham, 2016).

Figure 2. Correlation between the proportions of uses of verbs of the three languages covered by the other two languages



Among other things, Figure 2 shows which verbs in each language are poorly covered by the sight verbs of the other two languages: these verbs lie in the bottom left corners of the graphs. These verbs are likely to have meaning extensions into other domains, such as the Russian verb *rassmatrivat'* 'examine' discussed in section 2. The detailed analysis of these meaning extensions is, however, beyond the scope of this paper.

4. Mutual pairwise correspondences

As the first take at the semantic similarity between the verbs, for each pair of languages, I identified the pairs of verbs which show the highest degree of correspondence to each other. This degree was measured using Log-likelihood score, which is used, among other fields, in collocation analysis to express the degree of attraction between collocates, i.e., words that frequently co-occur together in text (Evert, 2009; Su et al., 2024). Log-likelihood score (Dunning 1993) is based on the comparison of the observed and expected values for a contingency table with the cooccurrence data of the two words. In case of correspondences between sight verbs, such a contingency table was created for each pair of verbs, as exemplified in Table 5 for BG *gledam* 'look' and PL *patrzeć* 'look'.

Table 5. Observed and expected values for the correspondences between BG *gledam* 'look' and PL *patrzeć* 'look'

	Observed values		Totals	Expected values	
	gledam	The other BG sight verbs		*gledam*	The other BG sight verbs
patrzeć	19029	15867	34896	4366	30529
The other PL sight verbs	27372	308519	335891	42034	293857
Totals	46401	324386	370787		

To calculate the expected values, shown in the right part of Table 5, for each cell, the total by row is multiplied by the total by column and divided by the grand total. Then the log-likelihood score is calculated using the formula in (9), see (Evert 2009).

(9) log-likelihood =

$$2\sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

This way the tables with observed and expected values were constructed and the log-likelihood score was calculated for all the pairs of verbs in the three pairs of languages. Then the best matching verb was found for each of the verbs, see Tables 8-10 in the Appendix[7]. Basic sight verbs in these tables are given in bold, verbs that occur more than once are underlined.

The verbs that occur in the tables more than once happen to have the highest log-likelihood score for more than one verb in another language. This is inevitable, because in all the three pairs, the list of verbs for one of the languages has fewer verbs than that for the other. However, some verbs recur also in the languages with larger lists (Russian in both pairs, Polish as compared to Bulgarian).[8] This suggests that these recurring verbs have a higher degree of semantic generality, since their range of contexts is shared between at least two verbs in the other language. This applies, in particular, to the basic verbs, which belong to the recurring verbs in almost all the cases.

In all the three tables, the basic verbs of sight, especially ‘see’-verbs, as well as the verbs ‘notice’ and ‘observe’, are among the highest-ranking pairs. They are also among the most frequent verbs in the three languages. This raises the question of whether there is a correlation between the verbs’ frequency and their degree of correspondence, given that they are already established as a semantically best-matching pair.

Theoretically, two alternative hypotheses can be put forward concerning the relation between frequency and mutual correspondence between verbs of the same semantic class in a language. One possibility is that more frequent verbs, which are also likely to be more semantically general, are more similar to each other across languages, whereas less frequent members of the same class exhibit more irregularity and carve the semantic space in more idiosyncratic ways. Alternatively, the core members

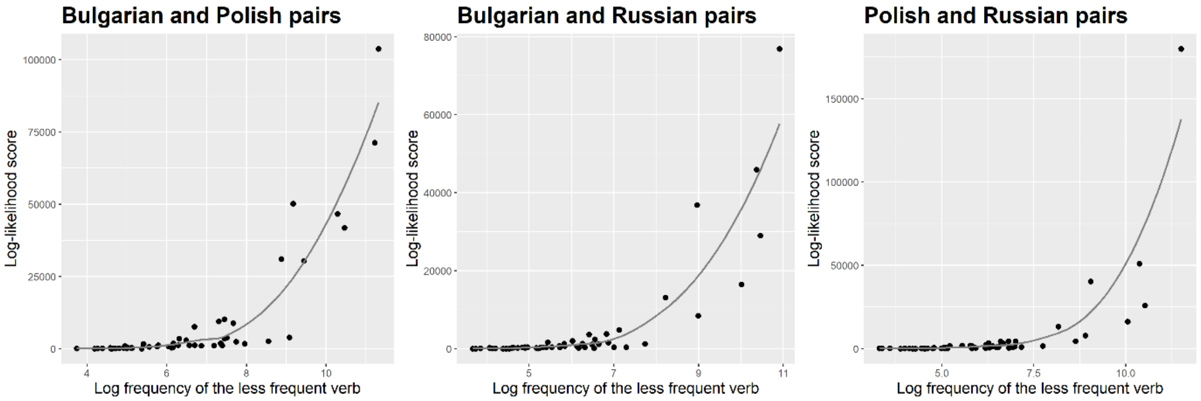
[7] In the first version of the study, the pairs of verbs with the highest degree of correspondence were identified using Dice-coefficient, which is another widely used collocation measure. The tables with pairs based on Dice-coefficient are available at REMOVEDFORANONYMYTY. With a few exceptions, the two measures yielded the same pairs of verbs.

[8] Technically, the pairs with the highest log-likelihood scores were found for the language which had a larger number of verbs, e.g., for each Russian verb a Bulgarian counterpart was found. Then, I checked which verbs of the second language, in this example Bulgarian, are absent from the already found pairs and identified the Russian pairs for them.

of the class may be more prone to semantic extensions, which need not be the same across languages, while the more specific meanings of less frequent verbs might more neatly correspond to each other.

To test for the presence and the sign of the correlation, I used Spearman rank correlation test, which is a non-parametric test suitable for variables which are not normally distributed. For all the pairs of verbs, a significant very strong positive correlation is observed between the log-likelihood scores of the pairs of the best matching verbs and the log frequency of the less frequent verb of the pair (ρ = 0.92, *p* < 0.0001 for Bulgarian and Polish; ρ = 0.85, *p* < 0.0001 for Bulgarian and Russian; ρ = 0.85, *p* < 0.0001 for Polish and Russian). Figure 2 shows these correlations for the three pairs of languages. Each point corresponds to one pair of verbs, the curve summarizes the trend.

Figure 2. Correlations between the verbs' frequency and the pair's log-likelihood score



The verbs in best-matching pairs can be of very different frequencies. To check whether the same correlations are observed when the frequency ratio is more balanced, I additionally tested only the pairs of verbs with comparable frequencies (where the frequency of one verb is more than 0.5 and less than 1.5 of the frequency of the other). Again, strong or very strong positive correlations were found between the log-likelihood scores and log mean frequencies of the verbs in a pair for all the three pairs of languages (ρ = 0.97, *p* < 0.0001 for Bulgarian and Polish, 12 pairs of verbs; ρ = 0.74, *p* ≈ 0.001 for Bulgarian and Russian, 17 pairs; ρ = 0.97, *p* < 0.0001 for Polish and Russian, 15 pairs).[9]

These correlations favor the hypothesis of the positive correlation between the verbs' frequency and the strength of the mutual correspondence. This correlation may suggest that the mutual correspondence between more frequent verbs is not so much affected by differences in the minor patterns resulting from language-specific semantic extensions, whereas for less frequent verbs, even minor discrepancies in usage patterns can result in a considerably lower degree of mutual correspondence. Consider again the correspondences between the Polish verb *wyjrzeć* and the Russian

[9] In the previous version of the study, the same correlations were tested and found significant using the standardized Dice-coefficients. As Dice-coefficient is dependent on the verbs' frequency, the observed Dice-coefficients were expressed as z-scores based on the distribution of randomly generated Dice-coefficients.

verb *vygljanut'*, discussed in section 2. Although these two verbs seem to match very closely both derivationally and semantically, their mutual correspondence, at least in the context of looking out of a window, is affected by the fact that Russian more often uses basic verbs of sight for this type of context. In the next section, I look at the extent to which the verbs under analysis correspond to basic sight verbs and how this correspondence can be interpreted in terms of semantic generality.

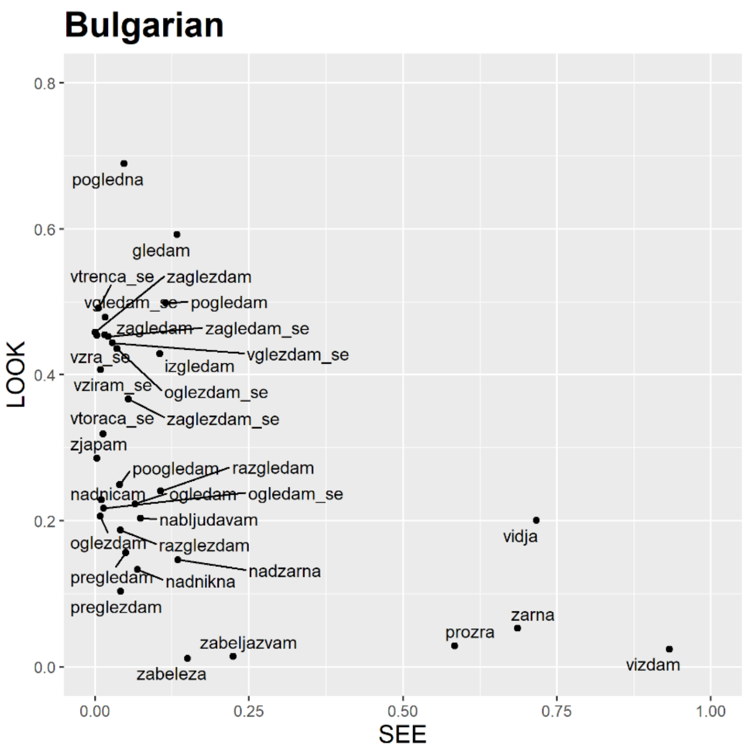
5. Similarity to basic sight verbs

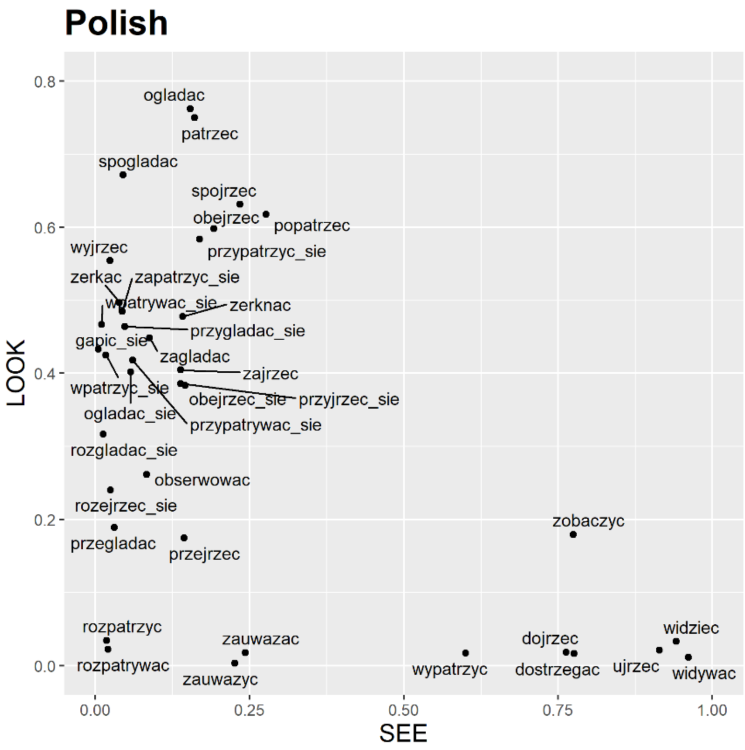
As mentioned in the previous section, for some lower-frequency sight verbs, one of the basic verbs of sight serves as the best-matching verb. The high frequency of correspondence of more semantically specific verbs to basic verbs of sight is not surprising, since basic verbs of sight denote sight in its most general form and are stylistically neutral. Some verbs of sight can be more similar to ‘look’-verbs and some, to ‘see’-verbs. In this section, I explore this question by using basic verbs of sight as two poles of the sight domain and as an ideal representation of the ‘look’ vs. ‘see’ distinction. Assessing the similarity of various sight verbs to these two poles can show whether this distinction is relevant for the structuring of the sight domain as a whole or only valid for basic verbs.

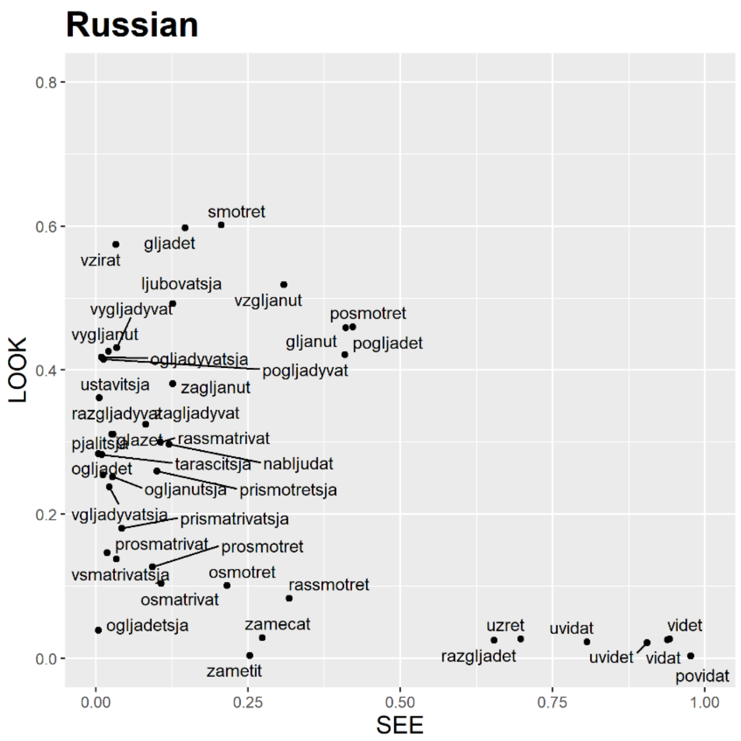
For each of the verbs under analysis, I look at their similarity to basic verbs in the other two languages, based on the frequency of their correspondence. For instance, for each Bulgarian verb, I calculated the proportions of uses where it corresponds to the basic ‘see’-verbs in Polish and in Russian and the mean of these proportions, as well as the proportion of uses where it corresponds to the basic ‘look’-verbs in these languages, and the mean of the two[10]. These mean proportions for each of the three languages are visualized in a two-dimensional graph, see Figure 3. The x-axes of the graphs correspond to the mean proportions of correspondences with basic ‘see’-verbs and the y-axes, with ‘look’-verbs. The verbs that are close to the origin, i.e., are located in the lower left corner of the graph, rarely correspond to either basic ‘see’-verbs or ‘look’-verbs.

Figure 3. Proportions of correspondences to basic verbs

[10] This approach may have its drawbacks, because basic verbs of sight in each of the languages may have idiosyncratic patterns of use and different semantic ranges. This problem is somewhat alleviated by using two languages as the basis of comparison. A more language-neutral estimate can arguably be obtained by using a larger number of languages for comparison, as the idiosyncratic properties of individual verbs will be more levelled out.


Bulgarian
LOOK
SEE
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
pogledna
gledam
vtrenca_se
zaglezdam
vgledam_se
pogledam
zagledam
zagledam_se
vglezdam_se
vzra_se
izgledam
vziram_se
oglezdam_se
vtoraca_se
zaglezdam_se
zjapam
poogledam
razgledam
nadnicam
ogledam
ogledam_se
nabljudavam
oglezdam
razglezdam
nadzarna
pregledam
nadnikna
preglezdam
vidja
zarna
prozra
zabeljazvam
zabeleza
vizdam


Polish
LOOK
SEE
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
ogladac
patrzec
spogladac
spojrzec
obejrzec
popatrzec
przypatrzyc_sie
wyjrzec
zerkac
zapatrzyc_sie
wpatrywac_sie
zerknac
przygladac_sie
gapic_sie
zagladac
zajrzec
wpatrzyc_sie
obejrzec_sie
przyjrzec_sie
ogladac_sie
przypatrywac_sie
rozgladac_sie
obserwowac
rozejrzec_sie
zobaczyc
przegladac
przejrzec
rozpatrzyc
zauwazac
dojrzec
widziec
rozpatrywac
zauwazyc
wypatrzyc
dostrzegac
ujrzec
widywac



The general pattern shown by graphs in Figure 3 is that the verbs are mostly located along the axes, while the centres of the graphs remain empty. This means that the verbs gravitate toward the ‘see’-pole or the ‘look’-pole or to neither of them, but rarely share large parts of their contexts both with ‘see’-verbs and ‘look’-verbs. Thus, the distinction between ‘see’ and ‘look’ may indeed be viewed as structuring the domain of sight, including its non-basic members, with no intermediate zone between the two poles.

The most notable exceptions from this general pattern are the basic ‘see’-verbs BG *vidja* and PL *zobaczyć*, and the basic ‘look’-verb RU *posmotret'*, with several non-basic verbs located close to it (*gljanut'*, *pogljadet*, *vzgljanut'* ‘take a look’). The correspondences between these verbs are mainly observed in the contexts shown in (9) and (10).

(9) BG *no **vižte** kakvi sa tečenijata po šelfa sega.*
PL *ale **spójrz** na prądy wzdłuż szelfu teraz.*
RU *I **vzgljanite** teper' na tečenie vdol' berega.*
‘But look at the currents along the shelf now.’

(10) BG *Šte **vidja** ako možem da prosledim tezi xora.*
PL ***Zobaczę**, czy możemy wyśledzić tych ludzi.*
RU ***Posmotrju**, smožem li my otsledit' ètix ljudej.*
‘I'll see if we can't track these people down.’

In imperative contexts (9), Bulgarian systematically uses the verb *vidja* ‘see’ in contrast to the other two languages, and in first person future contexts (10), Russian is opposed to the other two languages, as it much more frequently uses the verb *posmotret'* ‘look’. Uses such as these are found in dialogical interaction and may be viewed as semantic extensions of the basic sight meaning. As such extensions are arguably more typical for basic than non-basic verbs, the basic ‘see’-verbs and ‘look’-verbs paradoxically show more similarity to each other than non-basic verbs lying close to one of these poles.

Figure 3 also shows that many more verbs tend toward the ‘look’-pole than the ‘see’-pole. This is reflected in the maximal proportions on the two axes. It is much closer to 1 in case of ‘see’-verbs, because they mostly correspond to themselves, whereas the basic ‘look’-verbs have many correspondences to non-basic verbs. This picture suggests that basic ‘look’-verbs are less diachronically stable and more to prone to renewal than basic ‘see’-verbs. This is also supported by the cognate relations between the basic ‘see’-verbs in the three languages and the etymological diversity of the basic ‘look’-verbs. As shown in Figure 3, some of the non-basic verbs, especially in Polish and Russian, are very close to the basic ‘look’-verbs, as *oglądać* ‘watch’ in Polish and *vzgljanut'* ‘take a look’ in Russian, denoting full perception and brief glance, respectively. In the next section, I show that these verbs are close to the basic ‘look’-verbs also in terms of frequency and the distribution of their correspondences to other verbs.

At the ‘see’-pole, there are also verbs that group close together with the basic verbs, such as the verbs *vidat'*, *uvidat'*, and *povidat'* ‘see’ in Russian. Although these verbs express perception as generally as basic verbs do, they have low frequency and are aspectually and/or stylistically restricted as compared to the basic verbs. Thus, they cannot be regarded as rivals to basic verbs, as the abovementioned verbs at the ‘look’-pole.

6. Evenness of the distribution of correspondences

One of the manifestations of semantic generality and contextual neutrality of basic verbs is that they can correspond to a wide range of non-basic verbs, irrespective of their specific semantic features. In the previous section, this was shown by plotting for the verbs of each of the languages the proportions of correspondences to basic verbs of the other two languages.

Non-basic verbs can also differ with respect to the distribution of correspondences to the verbs in the other languages. Some verbs mostly correspond to one or two verbs in another language, whereas the uses of some verbs are more evenly spread among the verbs of another language. For instance, the Bulgarian verb *zjapam* and the Polish verb *gapić się*, both meaning ‘stare’, while showing a high degree of mutual correspondence, have different distributions of correspondences. This is shown in Table 5, which contains five verbs (Polish and Bulgarian, respectively) with the highest frequencies of correspondence to the verbs *zjapam* and *gapić się*. The table shows raw correspondence frequencies and their proportions.

Table 5. Major correspondences of BG *zjapam* in Polish and of PL *gapić się* in Bulgarian

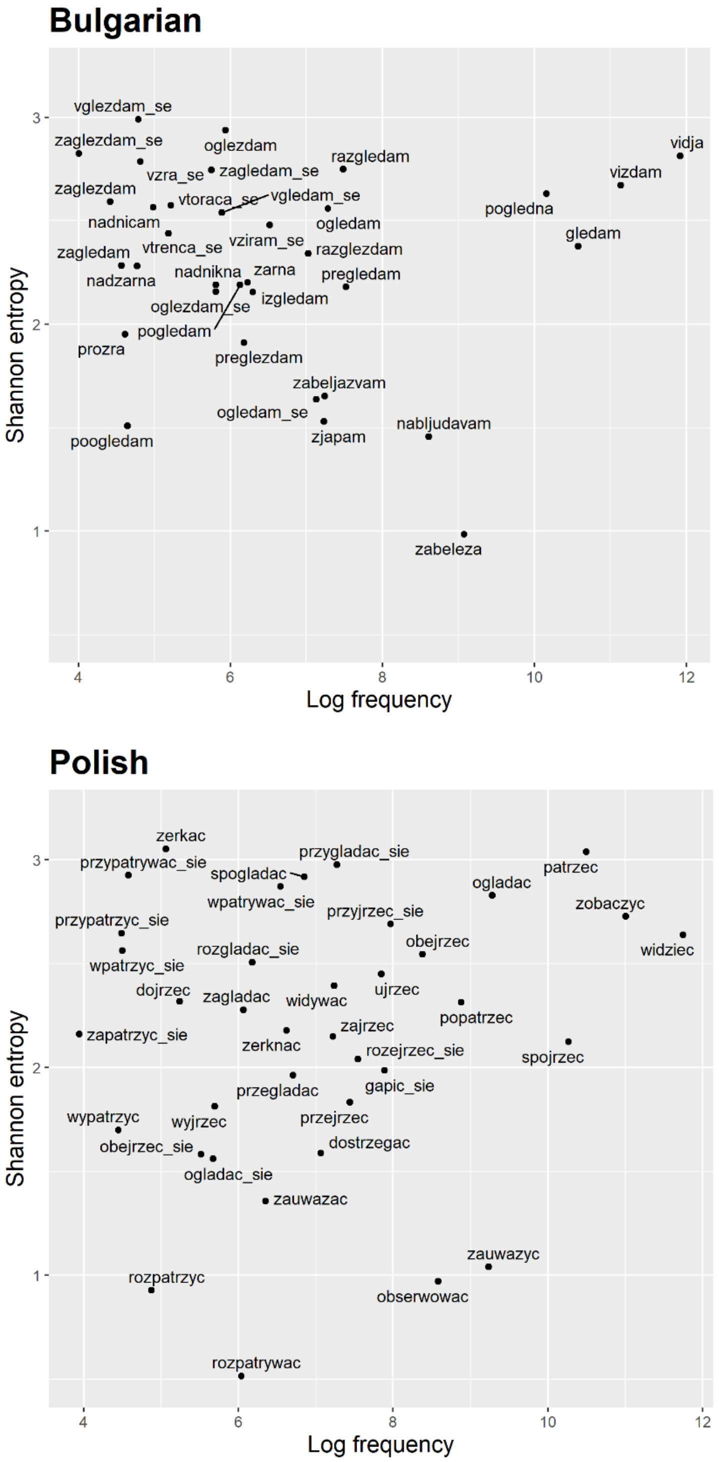
Polish verbs corresponding to BG *zjapam*			Bulgarian verbs corresponding to PL *gapić się*		
Verb	Raw frequency	Proportion	Verb	Raw frequency	Proportion
gapić się	1231	0.72	*gledam*	1451.12	0.45
patrzeć	296	0.17	*zjapam*	1231	0.39
wpatrywać się	42	0.02	*vziram se*	229	0.07
oglądać	37	0.02	*nabljudavam*	57	0.02
przyglądać się	33	0.02	*vtorača se*	54	0.02

The verb *zjapam* corresponds to the verb *gapić się* in more than two thirds of its uses, whereas the verb *gapić się* shares the majority of its uses between the verbs *gledam* and *zjapam*. This suggests that the Polish verb is more stylistically neutral than its Bulgarian counterpart.

The extent to which the distribution is concentrated within a small number of possible outcomes or more evenly spread among a larger number of outcomes can be measured with the Shannon entropy index, see, e.g., Stoll et al. (2017) for its application in linguistics. The higher the entropy of the distribution, the more evenly spread it is among the values. For example, for the verbs *zjapam* and *gapić się*, the Shannon entropy index based on the distribution of all the correspondences is 1.99 and 2.18, respectively.

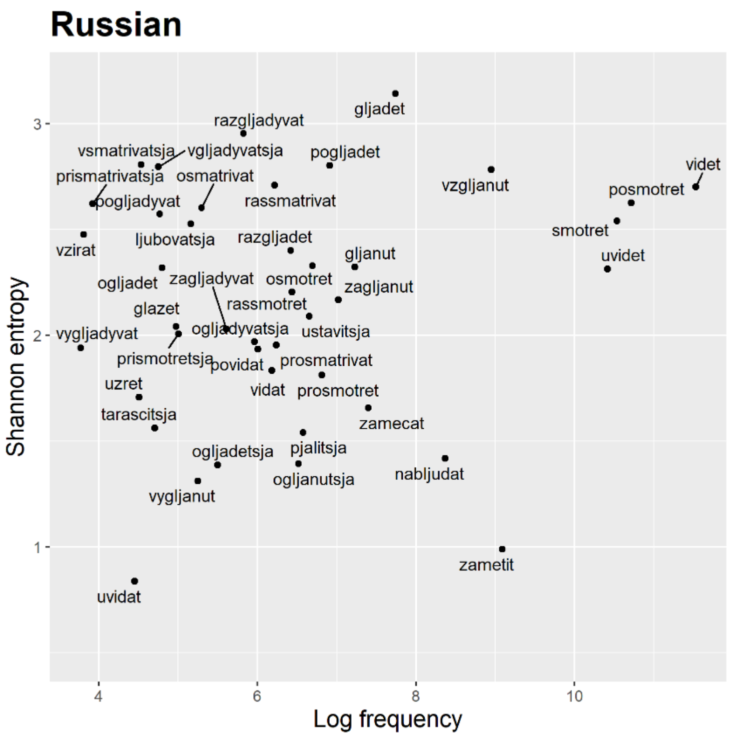
The proportions of correspondences to the basic verbs were analysed in the previous section and these proportions can considerably affect the entropy index. For these reasons, only the correspondences to non-basic verbs are taken into account in this section. For instance, for each Bulgarian verb (including the basic ones), the Shannon entropy index was calculated based on the distribution of uses for all the non-basic Polish and Russian verbs (using the R package *entropy* (Hauser & Strimmer, 2009)). In Figure 5, the values of Shannon entropy index for Bulgarian, Polish, and Russian verbs are plotted against the logarithm of the verbs' frequencies (the mean sum of correspondences to all the verbs in the two languages).

Figure 5. The entropy of the correspondences to non-basic verbs and verbs' frequency



Bulgarian

Polish



In Figure 5, the basic sight verbs are found in the top right part of the graphs: they have both high frequency and high values of entropy, which means that their correspondences to non-basic verbs are considerably spread out. Verbs close to the basic verbs in terms of frequency and entropy may be regarded as their (potential) rivals. Such verbs are especially numerous in Polish, e.g., *oglądać* and *popatrzeć*, testifying to a less clear-cut distinction between basic and non-basic verbs in this language. In Russian, the verb *vzgljanut'* 'take a quick look' is close to the area of the basic verbs. All these Polish and Russian verbs also show a high proportion of correspondences to basic 'look'-verbs, as shown in Figure 4 in section 5. In Bulgarian, there are no verbs as close to the basic verbs in terms of frequency and entropy. For all the three languages, a significant moderate to strong positive correlation is observed between entropy of correspondences to non-basic verbs and the proportion of correspondences to 'look'-verbs (Pearson's product moment correlation; r(32) = 0.42, $p \approx 0.01$ for Bulgarian; r(35) = 0.55, $p < 0.001$ for Polish; r(40) = 0.39, $p \approx 0.01$ for Russian). This again suggests that there is more rivalry and potential for lexical renewal among 'look'-verbs as compared to 'see'-verbs. No significant correlation was found between frequency and entropy.

The top left parts are densely populated in all three graphs. Many of the verbs found here denote focused or thorough visual perception and their correspondences are considerably spread between several verbs of this group, suggesting less clear semantic distinctions between them. As an illustration, Table 6 shows the correspondence frequencies for several Polish and Russian verbs from this group and their entropy values based on the correspondences to non-basic verbs (cells with perfective verbs are in grey).

Table 6. Correspondence frequencies and entropy values for several Polish and Russian verbs of focused and thorough perception

Verb	*przyjrzeć się*	*przyglądać się*	*wpatrywać się*	*wpatrzyć się*	Entropy of Russian verbs
rassmotret'	240.57	7	0	0	2.23
razgljadet'	131.52	4	0	0	2.41
rassmatrivat'	26	90	12	2	2.77
razgljadyvat'	20	74	20	1	3.00
prismotret'sja	99.99	6	1	0	2.01
ustavit'sja	8	16	59	16	2.18
vgljadyvat'sja	5	12	35	6	2.85
vsmatrivat'sja	1	13	31	1	2.84
prismatrivat'sja	8	17	0	0	2.69
Entropy of Polish verbs	2.70	2.98	2.87	2.56	

Although the distributions shown in Table 6, as viewed per rows and per columns, are far from uniform, they are typically not clearly skewed in favour of just one or two verbs, either. One of the reasons for this is the existence of synonymous Russian verbs with the roots *gljad*- and *smotr*- (similar pairs can also be found in Polish for the roots *patr*- and *gląd*-). However, other verbs also contribute to the fuzziness of the correspondences in this group. In particular, the Polish verbs with the prefix *przy*- most frequently correspond to the Russian verbs with the prefix *raz*-, but they also share some contexts with derivationally cognate verbs with the prefix *pri*-. For the verb *wpatrywać się*, the highest frequency of correspondence is observed with the verb *ustavit'sja*, but there are also many contexts it shares with other verbs in Table 6. Interestingly, imperfective verbs in Table 6 have higher entropy values than their perfective counterparts. Observationally, this generalization holds for the imperfective and perfective under analysis in general: in Figure 5, imperfective verbs typically have higher values than their perfective counterparts. This means that imperfectives tend to have a more evenly spread of correspondences to other verbs than perfectives. However, the statistical validation and interpretation of this observation will remain a matter of future research.

Naturally, there is no clear-cut distinction between the groups of verbs outlined above based on the graphs in Figure 5. Verbs with higher entropy values and higher frequency are expected to have a wider range of contexts and be less semantically specific, thus being closer to basic verbs in terms of semantic generality and contextual neutrality.

The verbs with lower entropy values found in the bottom parts of the graphs vary in frequency but they all more or less closely correspond to one verb in each of the remaining languages, as in cases of the verbs *nabljudavam*, *obserwować* and *nabljudat'* or *ogledam se*, *obejrzeć się*, *ogljadet'sja* and *ogljanut'sja*. These verbs also mostly have low proportions of correspondences to basic verbs, as shown in Figure 4. Thus, verbs from the lower part of Figure 5 are more semantically specific than those in the top parts of the graphs.

7. Semantic groupings in the sight domain

As shown in the previous sections, apart from the pairs of verbs with the highest degree of mutual similarity, measured here using log-likelihood score, there are more or less strong relations of mutual correspondence on other levels, whereby the majority of verbs under analysis show some degree of correspondence to a number of verbs in another language. Sight verbs within one language can also be more or less similar to one another. Visualization and analysis of these complex relations is the central topic of this section.

To explore the groupings of verbs both within and across languages, it is desirable to measure similarity between verbs within one language and the degree of correspondence between verbs in different languages in the same or in a similar way. In this case, the following approach was implemented. Similarity between verbs of the same language was measured, comparing the distributions of their correspondences to the verbs in the other two languages. For example, for BG *nadnikna* and *nadzărna* 'look over', I compared whether they have similar distributions of correspondences to all the Polish and Russian verbs under analysis. The degree of correspondence between verbs in two different languages was compared based on the distributions of their correspondences to the verbs in the remaining language. For instance, BG *nadnikna* 'look over' and RU *zagljanut* 'look in' were compared in terms of the similarity of their correspondences to the Polish verbs. Similarity between the two distributions was measured using the cosine similarity, which is widely used to compare documents on the basis of words they contain (Singhal, 2001); this was calculated using the function *cosine()* of the R package *lsa* (Wild, 2022). Then, cosine similarity was turned into distances between verbs. A fragment of the resulting distance matrix is given in Table 7.

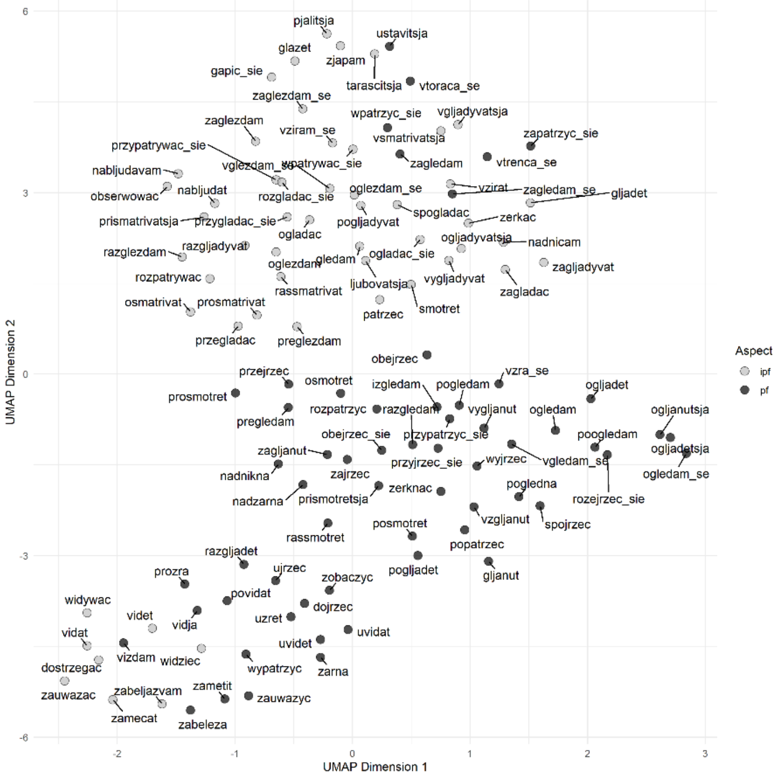
Table 7. A fragment of distance matrix with distances between the sight verbs

	gledam	*izgledam*	*nabljudavam*	*nadničam*	*nadnikna*	*nadzarna*	*ogledam*
zauważyć	0,98	0,98	0,95	0,98	0,98	0,97	0,99
zerkać	0,13	0,46	0,63	0,50	0,81	0,76	0,62
zerknąć	0,81	0,24	0,93	0,88	0,59	0,42	0,22
zobaczyć	0,78	0,47	0,90	0,93	0,82	0,66	0,55
glazet'	0,46	0,89	0,85	0,63	0,97	0,97	0,96
gljadet'	0,13	0,81	0,75	0,48	0,89	0,86	0,88
gljanut'	0,57	0,61	0,90	0,79	0,71	0,61	0,72
ljubovat'sja	0,05	0,72	0,60	0,58	0,94	0,90	0,88
nabljudat'	0,46	0,88	0,03	0,77	0,97	0,95	0,94

Lower distances indicate that the two verbs of different languages have similar distributions of correspondences to the verbs of the third language and are likely to be semantically more similar to each other, as, e.g., in case of BG *nabljudavam* and RU *nabljudat'*. Higher distances suggest less semantic similarity between verbs.

At the next stage, the pairwise distances between the 113 verbs under analysis were visualized using the UMAP (Uniform Manifold Approximation and Projection) algorithm, as implemented in the function *umap()* of the R package *umap* (Konopka 2023). This is one of the dimensionality-reduction techniques, which help represent distances between objects in a low number of dimensions and explore the grouping of these objects. Depending on the values set for the parameters of the UMAP visualization, the focus can be either more on the local structure, i.e., low-level groupings, or on the global structure of the data. Compared to the default settings of the function, I opted for the parameter values resulting in a more accurate rendering of the global structure with less tight clusters (n_neighbors was set to 50, and min_dist, to 0.5). The resulting visualization has a moderate correlation with the original distances (Spearman rank correlation, ρ ≈ 0.63), but the overall structure of the groupings remains the same irrespective of the model parameters.[11] Figure 6 shows the UMAP visualization of all the 113 verbs under analysis. The imperfective verbs are plotted by gray points, and the perfectives, by black points.

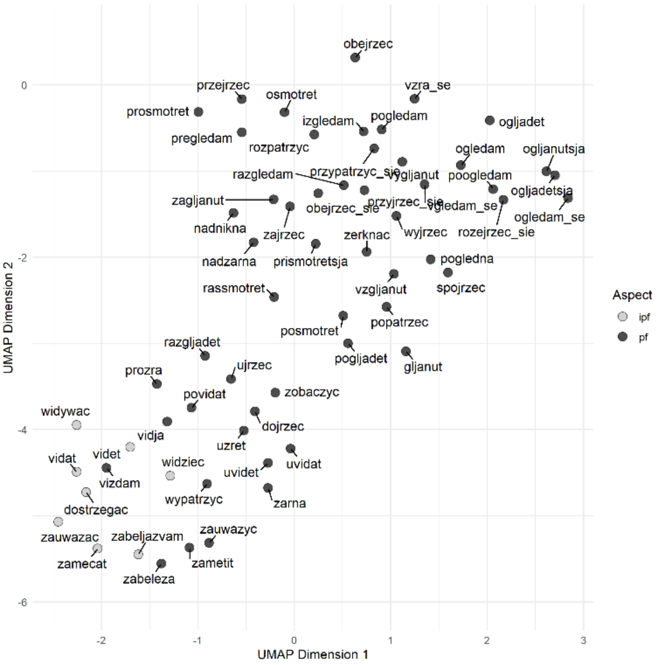
Figure 6. UMAP visualization for the sight verbs under analysis



[11] These groupings also largely recur when other visualization techniques are employed, such as Multidimensional scaling or hierarchical clustering.

The two semantic distinctions showing up in the larger groupings of verbs are aspect and the opposition of ‘see’- and ‘look’-verbs. These two distinctions result in three groupings: 1) imperfective and perfective ‘see’-verbs and other semantically similar verbs in the bottom left corner of the graph; 2) perfective ‘look’-verbs in the middle of the graph; 3) mostly imperfective ‘look’-verbs in the upper half of the graph. In Figures 7 and 8, the lower and the upper parts of the graph are represented separately to make their internal structure more perceptible.

Figure 7. Groupings of verbs in the UMAP visualization: lower part of the graph

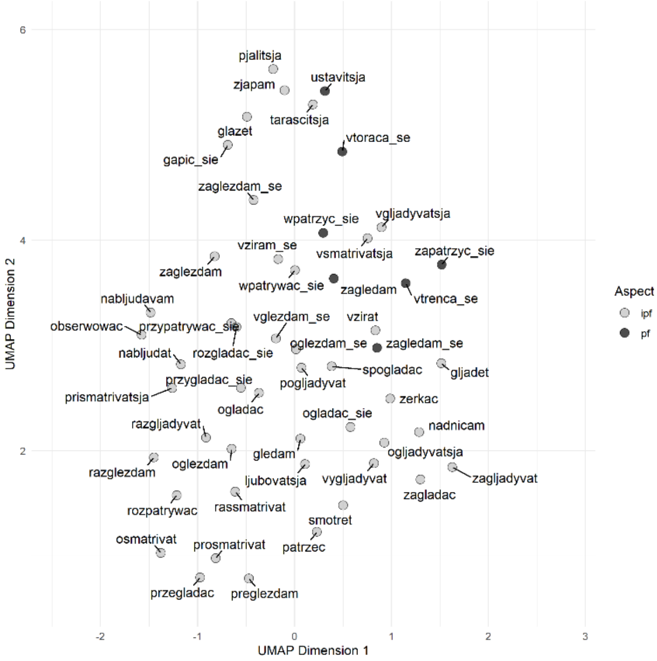


As mentioned above, in the bottom left corner of the graph, one finds basic ‘see’-verbs, such as BG *viždam*, as well as their more marginal quasisynonyms, such as RU *uvidat'*. This group also includes PL *wypatrzyć* and RU *razgljadet'*, which have the roots of ‘look’-verbs but denote the fact of perception specifying that it was accompanied by difficulties. At the very bottom of the graph, the ‘notice’ verbs of the three languages are situated, such as BG *zabeleža* and PL *zauważyć*. As shown in Figure 4 in section 5, they are indeed somewhat closer to ‘see’-verbs than to ‘look’-verbs. Note that this grouping is further subdivided in terms of aspect.

Although the perfective ‘look’-verbs in the second large grouping shown in Figure 7 do not fall into further distinct clusters, rather suggesting smooth transitions between the semantic types, some patterns may be identified here, too. Next to ‘see’-verbs, we find verbs of temporally limited perception, such as PL *popatrzeć* and RU *vzgljanut'*, together with basic perfective ‘look’-verbs (BG *pogledna*, PL *spojrzeć*), which also denote looking which is restricted in time. Above and to the right of this grouping, the perfective verbs of looking around are situated, such as BG *ogledam se* and PL *rozejrzeć się*. On the opposite side of this grouping, there is a small group of verbs which describe looking behind or over an obstacle, such as BG *nadzărna* and RU *zagljanut'*. The verbs found in the upper left part of Figure 7 convey the idea of thourough perception of an object, e.g., BG *pregledam*, PL *obejrzeć*, RU *osmotret'*.

Figure 8 shows the top part of the graph, where mostly imperfective 'look'-verbs are found.

Figure 8. Groupings of verbs in the UMAP visualization: upper part of the graph



At the very top of the graph, we find a relatively neat cluster of derogative verbs denoting perception of higher intensity, such as PL *gapić się* and RU *pjalit'sja*. Stylistically neutral verbs of this semantic type, such as BG *vziram se* and RU *vsmatrivat'sja*, are found further down the graph. In this group, among imperfective verbs, we find some inchoative perfective verbs that describe entry into this type of perception, such as BG *vtrenča se* and RU *ustavit'sja*. Thus, among these verbs, the type of perception overrides aspect, which otherwise is the major basis of the verbs' grouping.

The location of the remaining verbs shown in Figure 8 appears to be based on the duration and thoroughness of perception. On the left, we see verbs of attentive and thorough perception, such as BG *nabljudavam* and PL *przyglądać się*. Basic 'look'-verbs, such as PL *patrzyć* and RU *smotret'*, are found closer to the middle of the group. The verbs closer to the right denote a series of brief perception events, in particular related to change of position, e.g., BG *nadničam* 'look behind' and RU *ogljadyvat'sja*.

Based on these results, a tentative semantic classification of sight verbs in Slavic may be proposed, as shown in Table 8.

Table 8. Semantic groups of sight verbs as suggested by the UMAP-visualization

Semantic group	BG	PL	RU
Thorough perception verbs	*ogleždam*	*przejrzeć*	*rassmatrivat'*
Intense perception verbs	*zagledam se*	*gapić się*	*vsmatrivat'sja*
Temporally limited perception verbs,		*zerknąć*	*vzgljanut'*
including basic ‘look’-verbs (perfective)	*pogledna*	*spojrzeć*	*posmotret'*
Imperfective basic ‘look’-verbs	*gledam*	*patrzeć*	*smotret'*
Basic ‘see’-verbs	*viždam*	*widzieć*	*videt'*
and their (quasi-)synonyms	*zărna*	*wypatrzyć*	*povidat'*
Verbs denoting a particular spatial configuration:			
looking around	*poogledam*	*rozejrzeć się*	*ogljadet'sja*
looking behind or over an obstacle	*nadničam*	*zajrzeć*	*zagljanut'*

It should be stressed that the distinctions between some of these groups are not clear-cut; in particular, there is no clear borderline between thorough perception verbs, intense perception verbs, and basic ‘look’-verbs. Overall, the visualization testifies to the semantic complexity of the sight domain. Aspectual relations and the larger semantic groupings of verbs identified above are layered onto and stand out against the background of minor patterns of correspondence existing between verbs at various levels.

8. Conclusions

This paper contributes to the study of the semantics of sight verbs. It also demonstrates how parallel corpus data can provide a window into the complex structure of the sight domain. At the basis of the study is the premise that correspondences in parallel texts can be used to compare the verbs’ ranges of use and thereby establish the degree of their semantic similarity. The patterns of correspondence are not uniform across verbs – some verbs more neatly correspond to just one particular verb in another language, some have more or less frequent correspondences to a number of verbs in another language. By examining the multifaceted patterns of correspondence emerging in parallel texts, we gain insights into the processes of semantic extension, narrowing, and shifts that shape the relationships between verbs.

The degree of **mutual pairwise correspondence** between verbs helped identify the pairs of verbs closest to each other in terms of frequency of correspondence. Importantly, the degree of mutual correspondence in the best-matching pairs of verbs was found to positively correlate with the verbs’ frequency. This suggests the degree of correspondence between more frequent verbs is likely to be less sensitive to semantic shifts and extensions that affect one of them.

A large part of the paper focused on the distinction between basic and non-basic sight verbs. Basic verbs are the default and neutral means to describe sight in a most semantically general way and various non-basic verbs can be more or less similar to them. A straightforward measure of **similarity to basic verbs** is the proportion of uses of a sight verb in one language where it corresponds to the basic sight verbs in the other two languages. As basic sight verbs can be classified as either ‘look’-verbs or ‘see’-verbs, I calculated the proportions of correspondences to the basic ‘look’-verbs and to the basic ‘see’-verbs. Some sight verbs show semantic similarity to ‘see’-verbs, while much more verbs are similar to ‘look’-verbs, and some are far from either of the two poles. Still, there are no verbs intermediate between the two poles. Thus, the **distinction between ‘look’-verbs and ‘see’-verbs**, traditionally drawn for basic sight verbs, also plays an important role in structuring the domain as a whole.

The fact that there are more non-basic verbs similar to basic ‘look’-verbs than to basic ‘see’-verbs may be due to the fact that there are more manner and spatial configurations as well as the differences in assessment for controlled actions of directing attention rather than to the uncontrolled perception denoted by ‘see’-verbs. The presence of many verbs in the ‘look’-domain makes it more likely to be subject to lexical renewal and fluctuations in the verbs’ ranges of uses.

Semantic generality of basic sight verbs manifests in their ability to correspond to a wide range of diverse sight verbs. This results in the greater **evenness of the distribution of their correspondences** to the verbs in the other languages, in particular to non-basic verbs. This property was shown to be positively correlated with the proportion of correspondences to basic ‘look’-verbs. The third parameter that makes a verb similar to basic, and independent of the two already discussed, is frequency. Thus, verbs with a high proportion of correspondences to basic verbs of the other languages, with a more even distribution of uses across non-basic verbs, and with high frequency are closer to basic verbs of their own language and can be regarded as their potential substitutes.

The potential rivals of basic verbs can come from various semantic groups, e.g., verbs of brief glance, such as RU *vzgljanut'*, or complete perception, such as PL *oglądać*. Diachronically, a scenario of lexical renewal may be envisaged, whereby verbs with more specific meanings start to be used in a wider range of contexts and their specific semantic features can gradually fade. In some cases, they may reach a considerable degree of semantic generality and end up becoming new basic verbs. However, in most cases, they are likely to be soon ousted from their privileged position near basic verbs by other non-basic verbs undergoing semantic broadening.

Appendix

For each pair, the left column gives the language with a larger number of verbs. The basic sight verbs are given in bold.

Table 8. Bulgarian-Polish pairs of verbs with the highest degree of correspondence based on the log-likelihood score

Polish verbs	Aspect	Bulgarian verbs	Aspect	Log-likelihood score
widzieć	ipf	***viždam***	ipf	103774
zobaczyć	pf	***vidja***	pf	71370
zauważyć	pf	*zabeleža*	pf	50183
spojrzeć	pf	***pogledna***	pf	46671
patrzeć	ipf	***gledam***	ipf	41810
obserwować	ipf	*nabljudavam*	ipf	31123
oglądać	ipf	***gledam***	ipf	30410
gapić się	ipf	*zjapam*	ipf	10232
rozejrzeć się	pf	*ogledam se*	pf	9500
przejrzeć	pf	*pregledam*	pf	8854
rozpatrywać	ipf	*razgleždam*	ipf	7601
popatrzeć	pf	***pogledna***	pf	3911
rozejrzeć się	pf	*ogledam*	pf	3834
widywać	ipf	***viždam***	ipf	3539
przeglądać	ipf	*pregleždam*	ipf	3447
zauważać	ipf	*zabeljazvam*	ipf	2992
obejrzeć	pf	***gledam***	ipf	2589
przyjrzeć się	pf	*razgledam*	pf	2428
zajrzeć	pf	***pogledna***	pf	1919
rozglądać się	ipf	*ogleždam se*	ipf	1871
ujrzeć	pf	***vidja***	pf	1738
rozpatrzyć	pf	*razgledam*	pf	1681
wpatrywać się	ipf	*vziram se*	ipf	1255
obejrzeć	pf	*izgledam*	pf	1204
zajrzeć	pf	*nadnikna*	pf	1196
zerknać	pf	***pogledna***	pf	1190
przyglądać się	ipf	*nabljudavam*	ipf	1132
dostrzegać	ipf	***viždam***	ipf	1102
spoglądać	ipf	***gledam***	ipf	1050
rozejrzeć się	pf	*poogledam*	pf	937
rozglądać się	ipf	*ogleždam*	ipf	786
wyjrzeć	pf	***pogledna***	pf	771
przyjrzeć się	pf	*vgledam se*	pf	752
oglądać się	ipf	*ogleždam se*	ipf	676
zaglądać	ipf	*nadničam*	ipf	580
ujrzeć	pf	*zărna*	pf	490
popatrzeć	pf	*pogledam*	pf	358
gapić się	ipf	*vtoracha se*	pf	308
obejrzeć się	pf	***pogledna***	pf	301
wpatrywać się	ipf	*zagledam*	pf	290
wpatrywać się	ipf	*vtrenča se*	pf	264
przejrzeć	pf	*prozra*	pf	233

zajrzeć	pf	*nadzărna*	pf	198
zerkać	ipf	*nadničam*	ipf	151
wpatrzyć się	pf	*zagledam se*	pf	116
gapić się	ipf	*zagleždam*	ipf	103
zapatrzyć się	pf	*zagledam se*	pf	86
wypatrzyć	pf	*zabeleža*	pf	85
przypatrzyć się	pf	***pogledna***	pf	81
gapić się	ipf	*zagleždam se*	ipf	70
wpatrywać się	ipf	*vzra se*	pf	60
przypatrywać się	ipf	*vgleždam se*	ipf	50
dojrzeć	pf	***vidja***	pf	42

Table 9. Bulgarian-Russian pairs of verbs with the highest degree of correspondence based on the log-likelihood score

Russian verb	Aspect	Bulgarian verb	Aspect	Log-likelihood score
videt'	ipf	***viždam***	ipf	76869
smotret'	ipf	***gledam***	ipf	45881
zametit'	pf	*zabeleža*	pf	36847
uvidet'	pf	***vidja***	pf	29023
posmotret'	pf	***pogledna***	pf	16528
nabljudat'	ipf	*nabljudavam*	ipf	13164
vzgljanut'	pf	***pogledna***	pf	8487
zamechat'	ipf	*zabeljazvam*	ipf	4845
prosmotret'	pf	*pregledam*	pf	3759
ogljanut'sja	pf	*ogledam se*	pf	3635
pjalit'sja	ipf	*zjapam*	ipf	2417
prosmatrivat'	ipf	*pregleždam*	ipf	1981
ogljadet'sja	pf	*ogledam se*	pf	1687
osmotret'	pf	*pregledam*	pf	1569
zagljanut'	pf	*nadnikna*	pf	1342
rassmatrivat'	ipf	*razgleždam*	ipf	1317
gljadet'	ipf	***gledam***	pf	1256
ustavit'sja	pf	*zjapam*	ipf	1163
rassmotret'	pf	*razgledam*	pf	1080
ogljadyvat'sja	ipf	*ogleždam*	ipf	748
gljadet'	ipf	*zagledam se*	pf	707
posmotret'	pf	*pogledam*	pf	577
zagljadyvat'	ipf	*nadničam*	ipf	458
osmatrivat'	ipf	*pregleždam*	ipf	429
gljanut'	pf	***pogledna***	pf	428
pogljadet'	pf	***pogledna***	pf	410
glazet'	ipf	*zjapam*	ipf	395
razgljadyvat'	ipf	*razgleždam*	ipf	392
vygljanut'	pf	***pogledna***	pf	391
zagljanut'	pf	*nadzărna*	pf	385

ogljadet'	pf	*ogledam*	pf	368
ustavit'sja	pf	*vtrenča se*	pf	348
posmotret'	pf	*izgledam*	pf	346
ustavit'sja	pf	*vtorača se*	pf	340
vidat'	ipf	***viždam***	ipf	339
taraščit'sja	ipf	*zjapam*	ipf	285
vgljadyvat'sja	ipf	*vziram se*	ipf	253
prismotret'sja	pf	*vgledam se*	pf	248
vsmatrivat'sja	ipf	*vziram se*	ipf	241
ljubovat'sja	ipf	***gledam***	ipf	239
povidat'	pf	***vidja***	pf	230
vygljadyvat'	ipf	*nadničam*	ipf	200
smotret'	ipf	*ogleždam se*	ipf	187
razgljadet'	pf	*ogledam*	pf	165
ustavit'sja	pf	*vzra se*	pf	93
gljadet'	ipf	*zagledam*	pf	80
pjalit'sja	ipf	*zagleždam*	ipf	72
prismatrivat'sja	ipf	*vgleždam se*	ipf	57
uvidat'	pf	***vidja***	pf	55
ogljadet'sja	pf	*poogledam*	pf	54
vzirat'	pf	***gledam***	pf	48
pjalit'sja	ipf	*zagleždam se*	ipf	46
uzret'	pf	*zărna*	pf	44
pogljadyvat'	ipf	***gledam***	ipf	34
uvidet'	pf	*prozra*	pf	31

Table 10. Polish-Russian pairs of verbs with the highest degree of correspondence based on log-likelihood score

Russian verbs	Aspect	Polish verbs	Aspect	Log-likelihood score
videt'	ipf	***widzieć***	ipf	179899
uvidet'	pf	***zobaczyć***	pf	50973
zametit'	pf	*zauważyć*	pf	40333
smotret'	ipf	***patrzeć***	ipf	25800
posmotret'	pf	***spojrzeć***	pf	16148
nabljudat'	ipf	*obserwować*	ipf	13149
vzgljanut'	pf	***spojrzeć***	pf	7931
posmotret'	pf	*popatrzeć*	pf	4506
prosmotret'	pf	*przejrzeć*	pf	4472
zagljanut'	pf	*zajrzeć*	pf	4357
pjalit'sja	ipf	*gapić się*	ipf	4269
prosmatrivat'	ipf	*przeglądać*	ipf	3290
ustavit'sja	pf	*gapić się*	ipf	3188
ogljanut'sja	pf	*rozejrzeć się*	pf	3028
zamečat'	ipf	*zauważać*	ipf	2319
ogljadyvat'sja	ipf	*oglądać się*	ipf	1956

ogljadet'sja	pf	*rozejrzeć się*	pf	1912
zagljadyvat'	ipf	*zaglądać*	ipf	1821
vygljanut'	pf	*wyjrzeć*	pf	1759
rassmotret'	pf	*przyjrzeć się*	pf	1543
gljadet'	ipf	***patrzeć***	ipf	1541
ogljanut'sja	pf	*obejrzeć się*	pf	1354
videt'	ipf	*widywać*	ipf	1307
osmotret'	pf	*obejrzeć*	pf	988
zamečat'	ipf	*dostrzegać*	ipf	956
vzgljanut'	pf	*zerknać*	pf	920
gljanut'	pf	***spojrzeć***	pf	886
prismotret'sja	pf	*przyjrzeć się*	pf	745
razgljadet'	pf	*przyjrzeć się*	pf	677
taraščit'sja	ipf	*gapić się*	ipf	630
ogljadyvat'sja	ipf	*rozglądać się*	ipf	561
rassmatrivat'	ipf	*przyglądać się*	ipf	502
glazet'	ipf	*gapić się*	ipf	489
vidat'	ipf	***widzieć***	ipf	466
pogljadet'	pf	***spojrzeć***	pf	458
razgljadyvat'	ipf	*przyglądać się*	ipf	422
rassmotret'	pf	*rozpatrzyć*	pf	289
vgljadyvat'sja	ipf	*wpatrywać się*	ipf	274
pogljadyvat'	ipf	*spoglądać*	ipf	269
vsmatrivat'sja	ipf	*wpatrywać się*	ipf	256
pogljadyvat'	ipf	*zerkać*	ipf	247
rassmatrivat'	ipf	*rozpatrywać*	ipf	221
ogljadet'	pf	*rozejrzeć się*	pf	184
osmatrivat'	ipf	*oglądać*	ipf	130
uzret'	pf	*ujrzeć*	pf	125
uvidet'	pf	*dojrzeć*	pf	119
prismatrivat'sja	ipf	*przyglądać się*	ipf	114
uvidat'	pf	*ujrzeć*	pf	104
ljubovat'sja	ipf	*oglądać*	pf	100
ustavit'sja	pf	*wpatrzyć się*	pf	98
vygljadyvat'	pf	*wyjrzeć*	pf	75
povidat'	pf	***zobaczyć***	pf	72
ustavit'sja	pf	*zapatrzyć się*	pf	70
uvidet'	pf	*wypatrzyć*	pf	60
vzirat'	ipf	*wpatrywać się*	ipf	43
vgljadyvat'sja	ipf	*przypatrywać się*	ipf	38
prismotret'sja	pf	*przypatrzyć się*	pf	20

References

Dickey, Stephen. 2000. Parameters of Slavic Aspect. A cognitive approach. Stanford, CACSLI Publications.

Divjak, Dagmar. 2015. Exploring the grammar of perception. A case study using data from Russian. Functions of Language 22.1, 44–68.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Computational linguistics 19.1, 61–74.

Evert, Stefan. 2009. Corpora and collocations. In: Anke Lüdeling & Merja Kytö (Eds.), Volume 2: An International Handbook, 1212–1248. Berlin, New York: De Gruyter Mouton. DOI: 10.1515/9783110213881.2.1212

Geniusiené, Emma. 1987. The Typology of Reflexives. Empirical approaches to language typology 2. Berlin, New York, Amsterdam: Mouton de Gruyter.

Georgakopoulos, Th., E. Grossman, D. Nikolaev & S. Polis. 2021. Universal and macro-areal patterns in the lexicon: A case-study in the perception-cognition domain. Linguistic Typology, 26(2), 2021–2088.

Hausser, Jean, and Korbinian Strimmer. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J. Mach. Learn. Res., 10, 1469–1484. Available online: https://jmlr.csail.mit.edu/papers/v10/hausser09a.html

Hilpert, Martin; Correia Saavedra, David. 2017. Why are grammatical elements more evenly dispersed than lexical elements? Assessing the roles of text frequency and semantic generality. Corpora, 12(3), 369–392. DOI: 10.3366/cor.2017.0125

Kabacoff, R. I. 2011. R in Action: Data analysis and graphics with R. Shelter Island: Manning.

Knjazev, Y. P. 2007. Грамматическая семантика: Русский язык в типологической перспективе. Москва.

Konopka, T. 2023. _umap: Uniform Manifold Approximation and Projection_. R package version 0.2.10.0, <https://CRAN.R-project.org/package=umap>.

Levshina, Natalia. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. Folia Linguistica, 50(2), 507–542. DOI: 10.1515/flin-2016-0019.

Levshina, Natalia. 2017. Online film subtitles as a corpus: An n-gram approach. Corpora, 12(3), 311–338. DOI: 10.3366/cor.2017.0123.

Nesset, Tore, Laura A. Janda, Julia Kuznetsova, OX’PK Ljashevskaja, Anastasia Makarova, Svetlana Sokolova. 2008. Why poslushat′, but uslyshat′? Poljarnyj Vestnik 11. 38-46.

Nesset, Tore. 2010. Is the Choice of Prefix Arbitrary? Aspectual Prefixation and Russian Verbs of Perception. Slavic and East European Journal, 54(4), 666–689.

Norcliffe, Elisabeth, Asifa Majid. 2024. Verbs of perception: A quantitative typological study. Language, Vol. 100(1), 81-123.

Padučeva, E. V. 2004. *Dinamičeskie modeli v semantike leksiki.* Moscow: Jazyki slavjanskoj kul′tury.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rosen, A., M. Vavřín, A. J. Zasina. 2022. InterCorp, Release 15 of 11 November 2022. Institute of the Czech National Corpus, Charles University. Available at: http://www.korpus.cz

Rosen, Alexandr. 2023. The InterCorp Parallel Corpus with a Uniform Annotation for All Languages. Journal of Linguistics/Jazykovedný časopis, 74(1), 254–267.

San Roque, Lila, Elisabeth Norcliffe, Kobin H. Kendrick, Majid, Asifa et al. 2015. Vision verbs dominate in conversation across cultures. In: Aikhenvald, Alexandra Y., and Anne Storch (eds.), Perception and Cognition in Language and Culture, 45–72. Brill. DOI: 10.1163/9789004280205_004.

San Roque, Lila, Kendrick, Kobin H., Norcliffe, Elisabeth, Majid, Asifa. 2018. Universal meaning extensions of perception verbs are grounded in interaction. Cognitive Linguistics, 371-406. https://doi.org/10.1515/cog-2017-0034

Singhal, Amit. 2001. Modern Information Retrieval: A Brief Overview. IEEE Data Engineering Bulletin, 24(4), 35–43.

Stoll S., Mazara J., Bickel B. (2017) The acquisition of polysynthetic verb forms in Chintang // Fortescue M., Mithun M., Evans N. (Eds.) The Oxford Handbook of Polysynthesis. Oxford: Oxford University Press, 495–514.

Su, Q., Gu, C., & Liu, P. (2023). Association measures for collocation extraction. International Journal of Corpus Linguistics, 28(1), 59–86. https://doi.org/10.1075/ijcl.21056.su

Sweetser. 1990. From Etymology to Pragmatics. Metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. Journal of Child Language, 31(1), 61–99. https://doi.org/10.1017/S0305000903005881&#8203

Viberg, Åke. 1984. The verbs of perception: a typological study. *Linguistics* 21.1, 123–162.

Viberg, Åke. 2001. Verbs of perception. In: M. Haspelmath, E. König, W. Oesterreicher, & W. Raible (Eds.), Language Typology and Language Universals: An International Handbook, pp. 1294–1309. Berlin: Walter de Gruyter.

von Waldenfels, R. 2012. Aspect in the imperative across Slavic — a corpus driven pilot study. In: Atle Grønn & Anna Pazelskaja, eds., The Russian Verb. Oslo Studies in Language 4, 141–154.

Wälchli, B. 2016. Non-specific, specific and obscured perception verbs in Baltic languages. Baltic Linguistics, 7, 53–135.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wild, F. 2022. lsa: Latent Semantic Analysis. R package version 0.73.3, <https://CRAN.R-project.org/package=lsa>.