Proceedings of the

Fourth International Conference

Computational Linguistics in Bulgaria

25 – 26 June 2020
Sofia, Bulgaria

Syntactic and morphological features after verbs of perception: Bulgarian in Balkan context

Ekaterina Tarpomanova
Faculty of Slavic Studies
Sofia University Saint Kliment Ohridski
katya@slav.uni-sofia.bg

Abstract

The paper analyses the types of constructions that express a subordinate event after a verb of perception in the languages of the Balkan *Sprachbund*. The subordinate clauses that may follow a verb of perception are a result of common historical processes in Bulgarian, Albanian, Romanian and Greek: the substitution of infinitive by subjunctive and the neutralization of modal and declarative conjunctions after verbs of perception. Additionally, in Albanian and Romanian among the non-finite verbal forms gerund may be found after perception verbs. For the analysed syntactic structures in Bulgarian a corpus approach is further applied in order to support the linguistic analysis with quantitative data.

Keywords: perception verbs, syntactic structure, verb tense and aspect, Balkan languages

1. Introduction

Verbs of perception are a group of verbs whose semantics is related to the experience of one of the senses (traditionally recognized as vision, hearing, taste, smell, and touch, but also internal experiences such as feeling). In a 1984 paper Viberg presents a markedness hierarchy of the perception verbs based on a crosslinguistic study covering 50 languages, concluding that the sense modality hierarchy is the following: sight > hearing > other modalities (Viberg, 1984). As a semantic group, predicates of perception have similar argument structure involving an experiencer who receives the sensory information and a stimulus that prompts the sensory feeling. This study discusses the syntactic and morphological properties of the subordinate clauses or non-finite verb forms that follow the perception predicate expressing a second event the experiencer perceives in the languages of the Balkan linguistic area focusing on Bulgarian.

In many Indo-European languages verbs of perception may add either a non-finite verb form or a clause introduced by a subordinating conjunction. In English non-finite verbal forms that may follow a perception verb are bare infinitive and gerund. In this construction the object of the perception verb is obligatory and it is in fact the logical subject of the non-finite form:

(1a) *I heard him/her* ***sing***.

(1b) *I heard him/her* ***singing***.

The two non-finite forms differ in the manner they present the event by the viewpoint of the experiencer: the bare infinitive describes the event as a whole, i.e. the experiencer heard somebody singing from the very beginning to the end; the gerund describes the event in its progress, i.e. the experiencer started hearing the song while somebody has already begun singing.

On the other hand, perception verbs may be followed by a subordinate clause with a finite verb form that allows for temporal marking (depending on the tense of the verb in the main clause and the rules of tense agreement), thus situating the event of the subordinate clause to the temporal axis with respect to the event of the main clause:

(2a) *I saw (that) she* ***came****.*
(2b) *I saw that he* ***was bleeding****.*
(2c) *I saw that she* ***has made*** *a lot of records.*
(2d) *When I saw that he* ***had died****, I literally cried myself to sleep.*

In the Balkan languages there are several types of subordinate clauses that may follow verbs of perception, and, additionally, among the non-finite verbal forms gerund may occur in Albanian and Romanian. In what follows, the models that may be found after perception verbs will be analysed and illustrated with examples extracted from corpora available online: the Bulgarian National Corpus (BulNC), the Reference Corpus of Contemporary Romanian (CoRoLa), the Albanian National Corpus (ANC) and the Corpus of Modern Greek (CMG). In addition, for the models found in Bulgarian quantitative data obtained by the BulNC will be presented.

2. The corpora

The abovementioned corpora are used as a source of authentic language examples to confirm the occurrence of the different types of constructions after perception verbs in the languages of the Balkan *Sprachbund*. Additionally, the BulNC is used for the corpus-based approach applied for Bulgarian with the aim to find out how the Balkan feature described here is spread in the language which is the focus of this conference.

The Bulgarian National Corpus is developed at the Institute for Bulgarian Language (Bulgarian Academy of Sciences). It consists of a monolingual (Bulgarian) part and 47 parallel corpora containing altogether 5.2 billion words. The Bulgarian part includes about 1.2 billion words in over 240 000 text samples classified by style, domain and genre and supplied with rich metadata. The monolingual annotation consists in tokenization, sentence splitting, POS tagging, lemmatisation and morphological annotation. The BulNC is dynamic and is constantly enriched with new texts (Koeva et al.,2012).

The Reference Corpus of the Modern Romanian Language was launched in December 2017 by the Research Institute for Artificial Intelligence and the Institute of Computer Science at the Romanian Academy. The CoRoLa contains both written and oral parts. The written texts comprise 1 billion+ tokens and are distributed in an unbalanced way in several language styles (legal, administrative, scientific, journalistic, imaginative, memoirs, blogposts), in four domains (arts and culture, nature, society, science). The written texts are automatically sentence-split, tokenized, part-of-speech tagged, and lemmatized (Barbu Mititelu, Tufiş, Irimia, 2018).

The Albanian National Corpus is developed by a team of linguists from Saint Petersburg (Institute for Linguistic Studies of the Russian Academy of Sciences) and Moscow (the School of Linguistics at HSE). It contains two main subcorpora: Corpus of the modern literary Albanian (main corpus) and Corpus of early Albanian texts. The main corpus contains 31.12 million words, distributed into four styles: press (75.2%), fiction (10.3%), nonfiction (13.8%), poetry (0.7%). The corpus is supplied with a morphological annotation (Morozova and Rusakov, 2015).

The Corpus of Modern Greek is created at the Russian Academy of Sciences using the web interface of the Eastern Armenian National Corpus. The corpus comprises 35.7 million tokens. The main text variety is journalism, additionally there are fiction texts, both Greek and translated. The search engine allows for searching by language variety (dimotiki or katharevousa) and by orthography (monotinic or polytonic) (Kisilier and Arhangel'skij 2018).

3. Substitution of infinitive in the Balkan *Sprachbund*

The loss or avoidance of infinitive is one of the main features of the Balkan morphosyntax (Asenova, 2002: 141). Infinitive has been replaced by subjunctive or subjunctive-like constructions. In Romanian and Albanian subjunctive has a weak morphological marking, thus differing from indicative only in 3 p. sg. and pl. in Romanian and in 2 and 3 p. sg. in Albanian, while in Greek due to phonetical reasons subjunctive has coincided with indicative. Bulgarian as a Slavic language originally has no subjunctive. In the conditions of a weak or missing morphological marking, the main subjunctive marker in the Balkan languages is the conjunction Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să*, which in Greek and Albanian grammar is considered a particle (Asenova, 2002: 150).

The substitution of infinitive by subjunctive constructions created an opposition between two types of subordinate clauses mentioned in the early studies of the similarities between the Balkan languages (cf. Sandfeld, 1930: 175): modal-voluntative introduced by the conjunction Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să*, and declarative introduced by the conjunctions Bulg. *че*, Alb. *se*, *që*, Gr. *πως*, *ότι*, *που*, Rom. *că*, *dacă*, *de* (cf. Asenova, 2002: 149). In certain circumstances the opposition between the two types of subordinators may be neutralized and this is the case of the clauses following a perception verb in the main clause:

(3a) Bulg. *Видях го/я да идва.*

(3b) Bulg. *Видях го/я, че идва.*

'I saw him/her coming.'

The use of subjunctive constructions after verbs of perception involve some restrictions in tense and aspect. In all Balkan languages only present is allowed in the subordinate clause introduced by the conjunction Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să*, exept for Albanian, where imperfect is possible too. The use of perfect is allowed after a negative form of perception verbs, but in this case the modal meaning of the conjunction is preserved, that is why it is not taken into consideration in the study. In Bulgarian and Greek, which have the grammatical category of aspect, subjunctive construction is generally one of the contexts that favor the use of perfective, but despite this fact after perception verbs only imperfective is possible. The exclusive use of the imperfective is motivated by the relation between the events in the main and the subordinate clause, the former being a point on the continuous line of the latter (Bakker, 1970: 81).

4. Constructions after verbs of perception in the Balkan languages

Several models of constructions occuring after verbs of perception may be outlined in the Balkan languages, some of them are due to their common diachronic development, others are bilateral similarities or language-specific peculiarities.

4.1. Subjunctive construction

The subjunctive construction, as mentioned previously, is the substitute of infinitive and an important similarity between the Balkan languages, including the context discussed here. It is introduced by the modal subordinator Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să* 'to', but after verbs of preception the conjunction has lost its modal functions.

(4a) Bulg. *Видях го да се усмихва.* (BulNC) 'I saw him smiling.'

(4b) Rom. *... dar nici nu l-am văzut să facă nici un compromis mare.* (CoRoLa) '… but I never saw him making any big compromise.'

(4c) Alb. *Më pëlqen shumë kur e dëgjoj të flasë.* (ANC) 'I like it very much when I listen to her talking.'

(4d) Gr. *Την είδα να πετάει...* (CMG) 'I saw her falling…'

As compared to the infinitive, the finite verb forms in the subjunctive construction are additionally marked for person, number and tense, but due to the temporal and aspectual restrictions, they do not bear any rich grammatical information. The infinitive is preserved in Romanian and in the north dialect of Albanian (Gheg), but it cannot be used after verbs of perception, which proves the limitation of its functions.

4.2. Declarative constructions

The declarative constructions after perception verbs are typically fronted by the subordinator Bulg. *че*, Alb. *se*, Gr. *πως*, *ότι*, Rom. *că* 'that'. Aditionally, in Albanian and Greek in this position may occur the so-called universal relative (the term is originally used by Petya Asenova to denote the invariable pronoun or pronominal adverb in the Balkan languages used in colloquial speech instead of inflected relative pronouns, cf. Asenova, 1983) – Alb. *që*, Gr. *που*. As stated above, declarative constructions allow for different tenses in the subordinate clause. The examples below show some of the possibilities after aorist in the main clause – perfect in (5a), present in (5b), imperfect in (5c), and aorist in (5d):

(5a) Bulg. *Видях, че не е помръднал.* (BulNC) 'I saw that he hasn't moved.'

(5b) Rom. ... *am văzut că se mișcă o umbră...* (CoRoLa) 'I saw that a shadow is moving…'

(5c) Alb. *Befas pashë se makina po drejtohej nga bulevardi.* (ANC) 'Suddenly I saw that a car was coming from the boulevard.'

(5d) *Είδα ότι τα χρυσαφικά άρχισαν να τελειώνουν.* (CMG) 'I saw that we were running out of jewels.'

Another option for declarative construction after perception verbs in the Balkan languages is a subordinate clause introduced by the pronominal adverb Bulg. *как*, Alb. *(se) si*, Gr. *πώς*, Rom. *cum* 'how'. In some contexts the adverb preserves the semantics of manner, but it may also be subjected to desemantization and used with a generalized sense just to register a fact without necessarily focusing on the manner the event is performed. This double role is demonstrated with the following examples:

(6a) Bulg. *Видях как загина.* (BulNC) 'I saw how he died.'

(6b) Bulg. *Видях как в тях проблесна облекчение.* (BulNC) 'I saw how they calmed down.'

(7a) Alb. *Pashë si u pushkatua vëllai i këngëtares.* (ANC) 'I saw how the singer's brother was killed.'

(7b) Alb. ... *dhe unë pashë se si u largua duke marrë me vete shprehjen enigmatike të syve të saj.* (ANC) '… and I saw her walking away, taking with her the enigmatic expression of her eyes.'

(8a) Gr. *Χαίρομαι που είδα πώς γίνεται.* (CMG) 'I'm glad I saw how it may be done.'

(8b) Gr. *Έτρεξα πίσω τους, μα σαν φθασα στην άκρολιμνιά, είδα πώς ήταν τρεις.* (Πηνελόπη Δέλτα, "Τον καιρό του Βουλγαροκτόνου") 'I ran after them, but when I reached the seashore, I saw they were three.'

(9a) Rom. ... *am văzut cum se face vinul în Dobrogea.* (CoRoLa) 'I saw the way wine is made in Dobrogea.'

(9b) Rom. *Gata, mă, l-am văzut cum a plecat pe șosea!* (CoRoLa) 'It's done, I saw him leaving on the road!'

In the sentences given above, examples indexed with (a) indicate the manner of realization, while in the ones indexed with (b) the adverb of manner is synonymous with the declarative conjunction. Disambiguation may only be made by the context and in some contexts both readings are possible. For Greek it should be noticed that one of the declarative conjunctions, *πως*, has derived from the pronominal adverb of manner and in the modern language they can be distinguished only in the written variety by the accent put on the adverb (*πώς* vs. *πως*).

4.3. Gerund

Among the non-finite verb forms, only gerund may occur after perception verbs in Albanian and Romanian. Similarly to other languages that allow for non-finite verb forms after perception verbs, direct object in the main clause is obligatory referring to the logical subject of the event expressed by the gerund.

(10a) Alb. *Unë nuk të pashë duke u nisur, sepse ti kishe marrë udhë në orët e vona të natës dhe nuk doje të ma prishje gjumin.* (ANC) 'I didn't see you leaving, because you left late at night and you didn't want to wake me up.'

(10b) Rom. ... *am văzut venind spre mine un bătrân...* (CoRoLa) '… I saw an old man coming to me…'

In Bulgarian and Greek the gerund always refers to the subject, therefore if a gerund us used after a perception verb, it refers to its subject and not to its direct object.

(11a) Bulg. *Видях го, тръгвайки за важно интервю.* 'I saw him when I was going to an important interview.'

(11b) Gr. *Τον είδα γυρίζοντας απ'το φροντιστήριο.* 'I saw him when I was coming back from school.'

5. Quantitative data for Bulgarian

In Bulgarian there are three competing constructions following verbs of perception: subordinate clauses fronted by *да* 'to', *че* 'that', and *как* 'how'. In this section, the frequency of their use and some characteristics of their syntactic structure is examined based on data of the Bulgarian National Corpus.

The rate of occurrence of the three subordinators has been surveyed after three basic perception verbs: 'see', 'hear' and 'feel' in 1 p. sg., aorist. The verb form is chosen as it is representative for the studied type of sentences and the results of the corpus search are focused and less noisy. Verbs for taste, smell and touch do not occur in that type of sentences (**I tasted him coming*). The results are presented in Table 1.

Verb / Conjunction	*че* 'that'	*как* 'how'	*да* 'to'
видях 'I saw'	5762	3191	1638
чух 'I heard'	3317	1347	1721
усетих 'I felt'	2413	1173	68

Table 1. Number of occurrences of the three conjunctions after perception verbs

The ratio between the conjunctions that introduce the different types of subordinate clauses after the three verb forms extracted from BulNC is displayed in Figure 1.

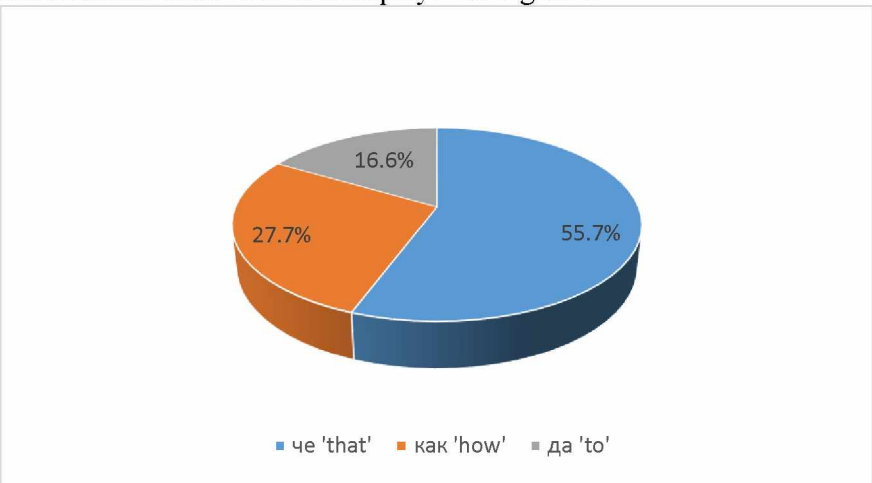

Figure 1. Ratio between the subordinators

The results of the corpus search show a clear preference for subordinate clauses introduced by 'that' in Bulgarian – more than a half as compared with the other two subordinators. The less used are 'to'-clauses – they slightly prevail over 'how'-clauses after the verb for hearing, but they concede considerably in the total result. The restricted use of the 'to'-clauses may be explained by the limited grammatical information that verbs in them may express.

Another issue that may be examined through corpus data is the syntactic structure of the main and the subordinate clause with respect to the argument realization. It is well known that in certain contexts the syntactic position of an argument may remain empty or may be filled by an argument of another predicate (Koeva, 2005: 37). This is the case of the perception verbs in many Indo-European languages whose object is in fact an argument of the predicate in the subordinate clause and its logical subject:

(12a) Eng. *I saw him/her come.*

(12b) Fr. *Je l'ai vu venir.*

In English and French this is the only syntactic structure possible, but in Bulgarian the logical subject of the subordinate predicate in all three constructions may be expressed either as an object of the main predicate (the verb of perception) or as a subject of the subordinate predicate. The argument may remain unexpressed only if the subordinate predicate is impersonal.

(13a) *Видях го/я да идва. / Видях той/тя да идва.*

(13b) *Видях го/я, че идва. / Видях, че той/тя идва.*

(13c) *Видях го/я как идва. / Видях как той/тя идва.*

'I saw him/her come.'

(14) *Видях да/че/как вали.* 'I saw it was raining.'

Nevertheless, the realization of the logical subject of the subordinate predicate as its grammatical subject is more typical for the declarative constructions, while as an object of the main predicate it is most often used in 'to'-constructions. This distribution is visible also by the corpus data presented in Table 2.

Verb / Conjunction	*че* 'that'	*как* 'how'	*да* 'to'
видях го 'I saw him'	226	301	683
чух го 'I heard him'	101	153	740
усетих го 'I felt him'	26	22	20

Table 2. Number of occurrences of the three conjunctions after perception verbs with explicit direct object

The corpus data show that there are no gender-specific differences in the realization of the logical subject of the subordinate predicate as a direct object of the main predicate. The search results with a feminine accusative pronoun in Table 3 display the preference for 'to'-constructions, except for the verb 'feel' whose limited occurrences are not statistically important.

Verb / Conjunction	*че* 'that'	*как* 'how'	*да* 'to'
видях я 'I saw her'	84	152	403
чух я 'I heard her'	33	80	356
усетих я 'I felt her'	15	12	4

Table 3. Number of occurrences of the three conjunctions after perception verbs with explicit direct object in feminine

Provided that the verb in Bulgarian is highly inflected the subject in a clause may be omitted. In sentences with a perception verb, if the argument of the subordinate predicate is realized as its subject, it can be omitted in the declarative 'that'- and 'how'-constructions, but never in the subjunctive 'to'-construction:

(15) *Видях, че идва. / Видях как идва. / *Видях да идва.* 'I saw (somebody) come.'

The fact that the subject cannot be omitted shows that the argument in the subjunctive construction is more naturally interpreted as an object of the main verb.

6. Conclusions and further directions

The study outlines several models of presenting a second (subordinate) event after verbs of perception in the languages of the Balkan *Sprachbund*. A common feature of the Balkan languages is the neutralization of the opposition between modal and declarative subordinators after perception verbs, but despite their semantic equivalence, the respective clauses they introduce differ in terms of possibilities for morphological marking of the verb and have some syntactic peculiarities related to the argument structure. The use of gerund that refers to the object of the main clause is a bilateral similarity between Romanian and Albanian, which is not shared by Bulgarian and Greek. Quantitative data for Bulgarian obtained by corpus search show that the most used model is the subordinate clause introduced by the declarative conjunction 'that' and that the realization of the logical subject of the subordinate predicate as an object of the main predicate is preferred in the 'to'-model. The study may be further enlarged by detecting translation equivalents of the models described here in parallel corpora.

Acknowledgements

This research was supported by the project *The Balkan languages as an emanation of the ethnical and cultural community of the Balkans (verb typology)*, financed by the Scientific Research Fund at the Ministry of Education and Science, contract ДН 20/9/11.12.2017.

References

Asenova, P. (1983). A propos des fonctions synatxiques des rélatifs absolus dans les langues balkaniques. In *Die slawische Sprachen 5. Referate des 2. Salzburger Slawistengesprächs "Probleme des Sprachkontakts"*, Teil 2, 5 – 12.

Asenova, P. (2002). *Balkansko ezikoznanie*. Veliko Tarnovo: Faber.

Bakker, W. F. (1970). The aspectual differences between the present and aorist subjunctives in Modern Greek. *Ελληνικά*, 23, 78 – 108.

Barbu Mititelu, V., D. Tufiș, E. Irimia. (2018). The Reference Corpus of Contemporary Romanian Language (CoRoLa). *Proceedings of the 11th Language Resources and Evaluation Conference – LREC'18*, Miyazaki, Japan, European Language Resources Association (ELRA), 1178–1185. Available at http://www.lrec-conf.org/proceedings/lrec2018/index.html.

Kisilier, M. L., T. A. Arhangel'skij. (2018). Korpusa grecheskogo iazyka: dostizheniia, celi i zadachi. *Indoevropeiskoe iazykoznanie i klassicheskaia filologiia*, XXII(1), 2018. P. 50 – 59.

Koeva, S. (2005). Argumenti – semantichni otnosheniya i sintaktichna realizatsiya. In Koeva, S., Ed., *Argumentna struktura. Problemi na prostoto i slozhnoto izrechenie.* Sofia: Semarsh, 25 – 42.

Koeva, S., I. Stoyanova, S. Leseva, Ts. Dimitrova, R. Dekova, E. Tarpomanova. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0 (1), 2012, 65 – 110.

Morozova, M., A. Rusakov. (2015). Albanian National Corpus: Composition, Text Processing and Corpus-Oriented Grammar Development. *Sprache und Kultur der Albaner. Zeitliche und räumliche Dimensionen. Akten der 5. Deutsch-albanischen kulturwissenschaftlichen Tagung (5.–8. Juni 2014, Buçimas bei Pogradec, Albanien)* / Hrsg. von B. Demiraj. Wiesbaden: Harrassowitz Verlag, 2015. (Albanische Forschungen, 37), 270 – 308.

Sandfeld, K. (1930). *Linguistique balkanique*. Problèmes et résultats. Paris.

Viberg, Å. (1984). The verbs of perception: a typological study. In Butterworth, B., Ed., *Explanations for language universals.* Berlin: Mouton, 123 – 162.

Corpora

ANC: Maria Morozova, Alexander Rusakov, Timofey Arkhangelskiy. Albanian National Corpus. (Available online at: albanian.web-corpora.net, accessed on 02.06.2020.)

BulNC: Bulgarian National Corpus, available at: http://search.dcl.bas.bg/, accessed on 02.06.2020

CoRoLa: Reference Corpus of Contemporary Romanian, available at: http://corola.racai.ro/, accessed on 02.06.2020

CMG: Corpus of Modern Greek, availab;e at: http://web-corpora.net/GreekCorpus/search/, accessed on 02.06.2020