Disambiguation in corpus of Modern Greek

Maxim Kisilier[1,2], Olga Nikolaenkova[1]
[1]Department of General Linguistics, Saint Petersburg University, Russia
[2]Institute for Linguistic Studies, Saint Petersburg, Russia
https://doi.org/10.36505/ExLing-2020/11/0027/000442

Abstract

Corpus of Modern Greek appeared in 2011. All texts are morphologically annotated. Due to certain peculiarities of Modern Greek morphology, the majority of forms has more than one grammatic interpretation. In this presentation we describe the types of homonyms which are found in the Corpus and discuss possible patterns for automatic disambiguation. At the end, we mention a number of problematic cases that cannot be resolved now or require manual approach.

Keywords: language corpus, ambiguity, Modern Greek, automatic disambiguation

General remarks

Corpus of Modern Greek (= CMG, http://web-corpora.net/GreekCorpus/search/?interface_language=en, access date 31.08.2020) was created in 2011 with the support of the "Corpus linguistics" program of the Russian Academy of Sciences. The size of CMG is over than 35.5 million tokens and some of its functions are absent from other corpora of Modern Greek (cf. Arkhangelskiy & Kisilier 2018). All texts in CMG are morphologically annotated by means of a digital grammatical dictionary and morphological analyzer (UniParser). The set of morphosyntactic values used for annotation coincide with basic grammatical categories (gender, number, case, tense, etc.).

Since the morphological annotation is an automatized process, each word has all possible analyses. Unlike Ancient Greek, flexions in Modern Greek often do not provide enough information to distinguish different forms (for example, αδελφ-ή 'sister' may be both nominative and accusative), and thus the percentage of ambiguous words and forms is high — according to Elizaveta Kuzmenko & Elmira Mustakimova (2015: 390) it is approximately 43%. It may grow with the further development of CMG and it is important to elaborate the mechanisms of disambiguation. Even now (when the corpus is not so large) manual disambiguation is not possible and at least the most typical cases should be disambiguated automatically.

Previous attempt of automatic disambiguation (Kuzmenko & Mustakimova 2015: 390) took into account only some definite articles, personal pronouns and certain forms of the verb 'to be'. Formally, these are the most frequent ambiguity examples in CMG, but from the point of view of Modern Greek morphology, they are less systematic than the homonymy of morphological

flexions. We believe that disambiguation in CMG requires a more systematic approach and, in this presentation, we intend to describe a number of most typical ambiguities and to discuss which of them do not require manual work. Most examples used in this paper are from CMG.

Lexical ambiguity

Lexical ambiguity, or homonyms is the best-known type of ambiguity and it is widespread in Modern Greek: βήμα — (a) 'step, pace', (b) 'tribune'. Such examples do not require any disambiguation at all as they do not affect morphological annotation.

Semilexical ambiguity

Semilexical homonyms usually belong to different morphological categories or classes:

(1) του (a) article in genitive
(b) personal clitic pronoun in genitive
(c) possessive pronoun

Their disambiguation could be based on syntax restrictions: (a) article always precedes NP, while (b) personal clitic pronoun is in front of a finite verb or after imperative/participle and (c) possessive pronoun follows either a noun or an adjective.

Sometimes situation, at first sight, looks more complicated:

(2) η **θεία** Αγάπη
ARTICLE aunt.NOUN Agapi
ARTICLE divine.ADJECTIVE.FEMININE love
'**aunt** Agapi (personal name)' or '**Divine** Love'

So far, (2) has no solution. Even a not fully proficient speaker of Modern Greek may get confused here. But let us take a look at (3) and (4) which illustrate the most typical usage of these homonyms:

(3) η **θεία** Ιουλία
ARTICLE aunt.NOUN Julia
'**aunt** Julia'

(4) η **θεία** λειτουργία
ARTICLE divine.ADJECTIVE.FEMININE liturgy
'**divine** liturgy'

Evidently, the noun (θεία/θείος 'aunt/uncle') is more commonly used both with a personal name or independently and the adjective is likely to be accompanied with a common name. If a number of semantic values is added to the grammatical dictionary in CMG, automatic disambiguation will be based on syntactic/combinatory restrictions. Certainly, some problematic situations, like (2), will not be resolved but their number will hardly exceed 2 or 3%.

Morphological ambiguity

This class includes several declension types where some flexions of different cases coincide, for example:

(5)

	SINGULAR		PLURAL
	NOMINATIVE	μητέρ-**α** 'mother'	μητέρ-**ες**
	ACCUSATIVE	μητέρ-**α**	μητέρ-**ες**

(6)

	SINGULAR		PLURAL
	NOMINATIVE		ψαράδ-**ες**
	GENITIVE	ψαρ-**ά** 'fishmen'	
	ACCUSATIVE	ψαρ-**ά**	ψαράδ-**ες**

Although (5) and (6) represent different declensions, disambiguation mechanism for all feminine and masculine nouns will be the same — the article will always indicate the right case. It is important to take into account that the article may be placed distantly:

(7)

η	**όμορφη**	γυναίκα	
ARTICLE	beautiful	woman	
'beautiful woman'			

(8)

η	**αγαπημένη**	**μου**	γυναίκα
ARTICLE	beloved	my	woman
'my beloved woman'			

(9)

η	**πολύ**	**αγαπημένη**	**μου**	γυναίκα
ARTICLE	very	beloved	my	woman
'my very much beloved woman'				

It is not very difficult to define the list of constituents which may separate the article from the noun (adjective/participle, possessive pronoun, few adverbs, etc.) even despite the fact that some types of constituents may be used more than once:

(10)

η	πολύ	**αγαπημένη**	μου	και	**ξεχωριστή**	γυναίκα
ARTICLE	very	beloved	my	and	exceptional	woman
'my very much beloved and special woman'						

Problems

One of the major challenges we face with neuter where the article does not help to distinguish nominative from accusative:

(11)

	SINGULAR	PLURAL
	NOMINATIVE	**το** λουλούδ-**ι** 'flower'	**τα** λουλούδ-**ια**
	ACCUSATIVE	**το** λουλούδ-**ι**	**τα** λουλούδ-**ια**

Modern Greek is a free word order language, that is why a syntactic regulation is not applicable here.

Another difficulty for automatic disambiguation are conjunctions ότι, που and πως which may be either complementizers or not. In (12), ότι is not a complementizer ('that') but an anaphoric pronoun (= ο τι):

(12) λέγε **ὁτι** θες
say what you.want
'say **whatever** you want'

However, only intonation or wider context helps to understand it. The same is relevant for another conjunction που:

(13) σε εσένα το λέω **που** μαζί τους συμφωνείς
to you it I.say who/that with them you.agree
'I say it to you **who** agree with them'

In (13), που does not refer to the adjacent verb but to pronoun εσένα.

(14) έλεγα <…> **πως** πολύ μου αρέσει,
I.said that/how much I like
πως είμαι περίεργος
that/how I.am curious

Both πως in (14) depend on the verb έλεγα despite the fact that the second πως immediately follows the verb αρέσει. Still without a wider context or intonation it is not clear whether πως means 'that' ('I said <…> **that** I like [it] very much, **that** I am curious [about it]') or 'how' ('I said <…> **how** much I like [it], **how** curious I am').

Certainly, there are some limitations in the use of complementizers, but the recent corpus-based analysis in (Kisilier 2020) clearly demonstrates that the system of complementizers is rapidly changing. Probably the best solution is to accept that in Modern Greek ὁτι, που and πως have multiple coexisting meanings which refer to the same word and are not homonymic.

Acknowledgements

This research was supported by the Russian Foundation for Basic Research (Project No 18-012-00607 "Corpus and fieldwork-based studies of sentential complements in the Balkan languages and dialects").

References

Arkhangelskiy, Timofey & Maxim Kisilier. 2018. Corpora of Modern Greek: achievements and goals (Корпуса греческого языка: достижения, цели и задачи). Indo-European Linguistics and Classical Philology 22(1), 50–59.

Kisilier, Maxim. 2020. Ὁτι, που and πως in Standard Modern Greek (Комплементайзеры ὁτι, που и πως в новогреческом языке). Indo-European Linguistics and Classical Philology 24(1). 554–577.

Kuzmenko, Elizaveta & Elmira Mustakimova. 2015. Automatic disambiguation in the corpora of Greek and Yiddish. In Computational linguistics and intellectual technologies (Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (2015) 1.) Annual International Conference "Dialogue". Vol. 1, 388–397. Moscow: Russian State University of Humanities.