Language and Power in Czech Corpora

Irene Elmerot [10000-0002-9809-8207]

Stockholm University, SE-106 91 Stockholm, Sweden
irel5167@student.su.se

Abstract. The author focuses on quantitatively examining the linguistic othering in printed media discourse in the Czech Republic, using the Czech National Corpus. The method used so far has been a corpus-based discourse analysis based on the adjectives preceding the keywords for each part of the project, now moving on to include reporting verbs. The theoretical starting point is that power relations in a society are reflected in that society's mainstream media, and that the language usage in these media contributes to the worldview of its recipients, in some cases even helps to construct it. Frequent but widely dispersed stereotypical and negative phrases and collocations are examples of a power language that may not be visible at once, but slowly enters the general discourse in a society. This project aims to survey these linguistic othering phrases in the Czech media discourse, as comprehensively as possible, and shed some light on their appearance over time.

Keywords: othering, discourse analysis, corpus linguistics, Czech

1 Introduction

This paper presents an ongoing project on how language relates to power and how othering is depicted in the language of a small but well-known country in the heart of the European continent.

2 Past project: Linguistic othering in Czech printed media

2.1 Roma vs. Gypsy – a short discourse and corpus analysis

The first article in this project was published in 2016 [6]. The subcorpus SYN of the Czech National Corpus (at that time about 5.170.696 lemmata¹ and 2.685.127.310

¹ For the sake of clarity, the following definitions are used: denomination: name of a group of people.
lemma (pl. lemmata): form of a word representing all forms of that word.
othering: the action of labelling someone who belongs to a different, often subordinate, social category with the purpose of exclusion from the sender's social category.

https://doi.org/10.26615/978-2-9701095-2-5_024
tokens [4]) was then used to analyse the Czech lemmata for Roma and Gypsy (Rom and Cikán) with their adjacent (position L−1) adjectives, to see what differences there were in the discourse of the most popular Czech printed media from 1989 to 2009, depending on which denomination had been used. A statistical analysis of the frequencies of adjectives adjacent to the two denominations was then performed. The main theory was a parallel to Masako Fidler’s idea [8] about finding a “more automatic mental representation” of these “others”. Most surprising of the results was that for both denominations, about a third of the adjectives in position L−1 were geographically related (as in “Romanian”, “Czech”, or “local”). The negative adjectives were about the double when adjacent to the lemma Gypsy compared to the lemma Roma. The neutral words were, on the other hand, almost the double for the lemma Roma compared to the lemma Gypsy. One adjective, “unadaptable” (nepřizpůsobivý), has become so popular in recent years as a collocation to the Roma people in the Czech Republic that it has created at least one article [13] and one book [12]. This adjective was found, but couldn’t really be classified as a collocation in this material. These are not very surprising results, but hereby confirmed for a “small” language by a large source material.

2.2 Linguistic othering of minorities in the Czech Republic (Follow-up)

In the follow-up [7] a similar method of analysing the frequency of adjectives adjacent to denominations of people was used, but this time the nouns for the minority groups Roma, Ukrainians and Vietnamese were analysed. The theory now focused more on the hypothesis that power relations in a society are reflected in that society’s mainstream media, and that these media’s language usage contributes strongly to the receivers’ worldview, in some cases even helps to construct it (cf. inter alia [8], [1] and [10]). The material was still the SYN series (now version 4) of the Czech National Corpus, and the time frame 1990–2014. The amount of lemmata had increased to 7,427,573, and tokens to 4,349,023,692 with this version [5]. The performed search returned 29,657 hits for the lemma Rom, 4,470 for Vietnamese (Vietnamec) and 5,335 for Ukrainian (Ukrajinec). Based on the previous research for these minority groups, the Roma were most likely to be linguistically othered in a similar way to the Vietnamese (who have been a large minority group in the Czech Republic since the 1960’s). It was not clear what would be found about the Ukrainians, who are a Slavic people like the Czechs, and have been immigrating in larger groups after 1989. The geographical words proved indeed to be much more frequent for Roma than for the other groups, e.g. of the total before the lemma Roma (29,657 absolute frequencies), “Czech” made up ten per cent. Before the lemma Vietnamese, there were a large frequency of the word “lonely” or “self” (samotný), and the word for “small” (malý) was also rather frequent – there was a double-check in context to make sure it didn’t mean “child”, but this did not seem to be the case. Both Roma and Vietnamese were on occasion considered “unadaptable”. The Ukrainians were found to be more positively

Also, “Roma” denominates all people of Roma descent, whether they call themselves Roma, Gypsy or something else. “The lemma Gypsy” or “the lemma Roma” then means the lemmata in the corpus searches or analyses.
depicted, but also often depicted as drunk (opilý). In total, the negative/positive ratio was 5.21 for the lemma Roma, 3.49 for Vietnamec and only 0.91 for Ukrajinec. A Pearson chi2 test has been performed on this, and the probability (P) value turned out to be 0.00, which means that we can reject the hypothesis that the variables are independent, i.e. that the observed differences in the sample data are systematic. Even with the neutral adjectives removed, the chi2 test shows the same result.

3  Future project: Linguistic power structures in Czech media

3.1  Gender structures shown in Czech media language (Work in progress)

During the autumn of 2017, the plan is to broaden the research scope of this project. Null hypothesis: There is no gender differentiation in the reporting verbs about women and men in the source material. Hypothesis 1: There is gender differentiation in the reporting verbs about women and men in the source material. To test this, the material will again be the most updated SYN series of the Czech National Corpus, and the keywords professional terms, like poslankyně (female member of parliament), úřednice (female office worker) and perhaps učitelka (female teacher), to compare theoretical professions with a more practical one where women are in majority. These keywords searches would then be filtered with reporting verbs like the Czech verbs for “claim”, “assert” or “establish”. The same searches and filtering would then be made for the male counterparts (MP, office worker and teacher) to see if the verbs differ in frequency, and what that may tell us about the linguistic othering structures created or added to by Czech media when it comes to working women of the middle and higher classes. One of the hypotheses will then be rejected. Previous research to form a basis will then be Baker [1], especially chapter 4 on how men and women are represented in a language, that is how mental images of the female professionals are created using “signifying practices and symbolic systems” [1, 89] in the language, and what the cumulative language usage tells us when such a large source material as the SYN version 5 corpus is analysed. Such corpora are likely to tell us what kind of language usage large numbers of people encounter regularly. Also research surveys such as Oates-Indruchová’s article [11] on the continued gender critique in the Czech Republic from 1948 onwards, and – for the background – such work as Rebecca Nash’s article [9] on gender scholars in the Czech Republic during the 1990s will be used.

3.2  Language structures: class, gender, minorities & the city vs. countryside cleft (planned project)

Since the aim is to turn this into a Ph.D. project in 2018, there is also the possibility of a larger project, where the corpus research could go wider in time and hopefully stretch back at least one century, since the Czech National Corpus is hoping to expand their amount of older corpora. The theory is not yet set in stone, but the previous research would include inter alia R. Čech’s paper on Language and ideology [2] and T.
Váňa’s Language power potential [14], as well as what M. Fidler and V. Cvrček are doing in their Needle in a Haystack project [10]. When it comes to the method to be used, the new collocation candidate function of the Czech National Corpus may come in handy here, to compare possible collocations to those words in the two corpora before and after 1989, and thereby extract expressions pointing towards the power structures and mental representations visible in the Czech media, as both Colson [3] and Fidler [8] mention. However, as Colson [3] points out, much caution must be taken when using such automated collocation extraction tools, as there may be collocations consisting of more than two words. Therefore, perhaps his CPR method may come in handy, or a similar method. The future will tell.

References