The Challenges of Data Collection in Contemporary Romanian Linguistic Research*

Anabella-Gloria Niculescu-Gorpin

The ”Iorgu Iordan – Al. Rosetti” Institute of Linguistics, the Romanian Academy

Linguistic hypotheses, whether related to synchronic or diachronic phenomena, are best tested on large data sets, i.e. on complex corpora that include a variety of language materials – spoken, written, video or audio. When setting out to develop such corpora, however, researchers often tend to put together material that can validate their assumptions about a particular issue, in other words they are mainly collecting data that are most likely to yield the results they expect to get out of them. Although potentially useful for linguistic research, such corpora have their disadvantages, the most important being that they are useful for only one or a limited number of researchers.

In my presentation I will try to address the challenges of data collection a Romanian linguist is currently facing. As my current research interest in present-day language change phenomena in Romanian requires a complex and extensive corpus, I soon discovered that there was no such corpus available for me to use. Faced with the daunting task of building a corpus myself with my own (limited) means, I started by looking at some of the major corpora available for other languages in order to identify some theoretical and methodological guidelines that could help me in my work.

I have looked at several corpora for other languages that I am hoping to (modestly) emulate, such as the British National Corpus or the National Corpus of Polish. In my presentation I will therefore give a brief overview of the lessons I have learned during that study, of the principles and methods that have been applied.

In the second part of my presentation I am going to make a comparative analysis of the current state of play in the domain of Romanian linguistic corpora. I must admit that the field is not a complete *tabula rasa*, as this would have made previous linguistic research impossible, as no language can be described or analysed in the absence of linguistic material. There are some corpora available, but they are generally purpose-built and can only serve the interests of a limited number of researchers. What is more, such corpora are difficult to use, as most of them are not available in an electronic format. So every time researchers are interested in studying a new topic, they have to start almost from scratch. The attempts that have been made so far to develop a diachronic and synchronic corpus of Romanian have not yielded any significant result. Due to the lack of coordination and cooperation among linguists, important resources in terms of funding and time have been wasted.

I will conclude by presenting the theoretical and practical solutions I have applied so far in my attempt to build a corpus and the plans I have to coagulate the efforts of Romanian linguists to eventually develop of National Corpus of Romanian. Of course, in a first stage I’m planning to put together the corpus that I need for studying contemporary language change in Romanian under the influence of English.

Word count: 496

*This work is supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-2480