Rafał L. Górski

Institute of Polish Language, Polish Academy of Sciences

REPRESENTATIVENESS OF A WRITTEN PART OF A POLISH GENERAL-REFERENCE CORPUS. PRIMARY NOTES

Abstract: The paper proposes a path towards solving the problem of representativeness of a large general-reference corpus of Polish. Its aim is not to propose an elaborated solution, but rather to narrow down the concept of representativeness and balance and show the methods leading to a final design of the corpus. Various notions of the concept of representativeness are discussed as well as their advantages, disadvantages and applicability. It is shown that representativeness and balance are often mutually exclusive requirements. The author concludes that the best solution for both theoretical and practical reasons is to represent the structure of readership in Poland.

Keywords: corpus design, general-reference corpus, representativeness.

The purpose of the following paper is to propose a path towards solving the problem of representativeness of the written component of a large general-reference corpus of Polish. I put aside the question of representativeness of the spoken component. My aim is not to propose an elaborated solution, but rather to narrow down the concept of representativeness and show the methods leading to a final design of the corpus. I shall also discuss the usefulness of several sources of data about readership in Poland, as well as an outline of proposed text typology.

Although all users expect corpora to be "representative", the concept itself seems to be far from being well established (Sinclair and Ball). There is no common consent what shall a corpus represent, i.e. what object it really represents. What is actually the object that corpus represents? Although we often say that it represents the particular language or a variety of language (e.g. Canadian English), it is of course not true. The language being an abstract phenomenon cannot be digitalized. A corpus consists of a number of texts, thus it represents the Saussurean ***parole;*** the corpus does not represent ***langue.*** This fact - regardless how self-evident it is - has consequences for setting the criteria of

representativeness. A corpus can only represent ***parole*** and so represent the population, the production, or the perception of texts.

On the one hand the concept of representativeness seems to be to some extent vague and the discussion of this topic pointless, on the other hand users expect the corpus they are provided with to be "representative and balanced." In fact the same query applied to different corpora of a given language may return different results depending on the design of the corpus.

Also considerable differences between general reference corpora of various languages can be observed.

As long as a written part is taken into consideration at least two approaches are possible. One is to represent the population of texts written (published) in a language in a specific span of time, the second is to represent the bulk of texts read by a linguistic community. Theoretically we can imagine five different approaches to the concept in question:

1. First imaginable approach is not to set any criteria, but rather concentrate on compiling a huge corpus on a random basis, which should cover a variety of texts. One can assume that, provided the corpus is really large, it would replicate the population of texts. This approach is rejected by the majority of corpus linguists because of its lack of any methodological basis. Note however that at the same time a large number of linguists (not necessarily devoted to corpus linguistics) use World Wide Web as a basis of linguistic research. The simplest, but very common way is to use Google to search for examples of usage. It is nothing else as this very approach.

2. Another solution is to set a number of text types and fill each text type with the same amount of running words. This approach was adopted for the so-called corpus of the frequency dictionary of Polish in the sixties (Kurcz ***et al.*** 1990). As far as I know nowadays no use is made of this approach. We can say that such a corpus is perfectly balanced, but it is not representative. This design of a corpus grants a marginal text type[1] the same representation as the ones which are most popular in means of production and perception, as, e.g., journalistic texts. It may be compared to a pool where the same number of members of the upper, middle and lower class are asked, although the quantity of each class is different.

3. The corpus may represent the population of texts[2]. This approach is methodologically very clear and seems to be the best solution. It is possible to

[1] By marginal I do not mean "of no social importance" but simply "not extensively read".

[2] In practice we can treat every single book or article as one item, but we can take into account also its length and/or circulation.

obtain an almost full bibliography of Polish print[3]. With all gaps in the bibliography we still can relatively easly define the population which we want to sample. Roughly this approach was adopted in compiling the "Brown Corpus family" (i.e. Brown, LOB, and Kolhapur corpus etc.). There is however only one concern about this approach, which makes it useless: namely the huge disproportion between the production of press and books. As I do not have data for Poland, I shall quote the data for Italy[4]: ca 93% of ninning words are printed in press and 7% in books. We can hardly expect Polish data to be very different. Even if such a corpus may be assumed as a representative one, it is not balanced. Still I consider this approach the best choice for historical corpora.

4. The corpus reflects the social stratification of the producers of the texts; this means applying to the written component the same criteria as to the spoken one. We should keep in mind however, that until quite lately there was a small number of those who published texts and a large majority of those who read it. In other words we are interested in a small group of text producers which are marginal in a linguistic community. However the situation changed recently. With blogs, Wikpedia and all what is called Web 2.0 the number of people who publish texts (and of course putting on Internet means publishing) has increased dramatically. To what extent these text are read - that is another question. This solution is methodologically not quite clear, also the its usefulness is questionable. I can hardly imagine arguments in fovour of this approach.

5. The corpus represents the perception of texts by a linguistic community. To my knowledge the first corpus which adopted this approach as the only criterion of representativeness is the Czech National Corpus[5] (Kralik & **Sulc** 2005). The design of the said corpus was based on a poll asking about what do the people read. The proportions of particular text types reflect the structure of readership. Somewhat oversimplifying - if statistically a member of a language community reads 12 novels and 6 manuals every year, the amount of belles-lettres in the corpus is twice as big as the amount of manuals. This approach - compared to representation of the population of texts - has some indubitable shortcomings. First of all it is much more difficult to define the structure of readership and the criteria are not as clear as sampling a closed list of books and newspapers. It is also much more expensive if the research has to be done from scratch. Nevertheless in my opinion the advantages outweigh the disadvantages.

[3] **Still, there is a number of printed books which are not captured by the Bibliographic Departament of the National Library, but it is rather a minor part of the whole production.**

[4] **Cf the homepage of CORIS/CORDIS corpus Design and implementation of a CORpus di Italiano Scritto http://corpora.dslo.unibo.it/coris_engDesign.html**

[5] **The criterion of readership was to some extent taken into account while disigning the British National Corpus.**

The corpus is somewhat better balanced (although this is not quite unproblematic). The main advantage is however of a more theoretical nature. Corpus linguistics view language as not merely mental but also - if not first of all - social phenomenon. Taking into account merely texts means neglecting the social context of the verbal communication.

As I said the approach which I suggest is not as unproblematic as one might see it at first glance.

First let us state: there is no average reader, but rather the corpus shall reflect the readership of the whole society. In fact I suggest to reflect in corpus the readership of persons who graduated institutions of tertiary education, because these people read much more then the rest of the society[6].

Fortunately every two years a detailed research of reading books in Poland is conducted. Even though the data do not meet all our needs, because they explore rather cultural aspects of reading than simply preferences in reading certain text types, they are useful for our purposes. As far as press is concerned, there are very precise figures relating to the number of people reading each newspaper and this fact may be taken into account. To my knowledge however nobody knows what is the amount of text in a given newspaper which is read by an average reader. This can make the task of stating the proportions of journalistic vs. non-journalistic texts difficult. Similar strictures apply to the readership of Internet.

Regardless what method we adopt there are two more problems. First we shall decide if the corpus should contain full texts or (as BNC) every text should be represented by a same number of running words. I am convinced that the latter is the only possible solution. As we aim at compiling a relatively large corpus on the one hand no single book can heavily influence the data and on the other hand we cannot "waste" the obtained material. Second: every corpus contains a category "other" or "miscellaneous". These texts as pamphlets, user guides, posters etc. which are not listed in the national bibliography, do not form a homogenic text type, and their readership cannot be easily compared to the readership of books and newspapers. In this case we have to decide arbitrarily the proportion of these texts in the corpus.

An elaboration of a clear text typology is a prerequisite of any research of reading which has to be done with the purpose of a corpus design. The question of such a typology is even more vague then the concept of representativeness. There is a considerable number of typologies, thus no consent is possible. Mostly the types are distinguished on a basis of extralinguistic factors. However having four different corpora of the Polish language we shall make a previous study to

[6] **I am aware of the fact, that this proposition is controversial.**

establish a typology based on purely linguistic and verifiable factors (cf. Biber 1988 and 1993; Utka 2004).

A classification of the text by topic shall be useful for lexicography, first to cover all major fields so as to obtain terrninology, as well as to help assigning terms to specific fields. The easiest method might be adopting a standard classification done by librarians. We should keep in mind however that these classifications are very detailed for academic writing but rather superficial for other domains of print.

Although we intend to elaborate a detailed conception of representativeness we consider implementing a device which will dynamically create corpora of different proportions eg. exactly the same as BNC, by searching not all but only a predifined list of texts, thus giving a comparable Polish corpus. It also might be useful to give the end-user a possibility of creating corpus of his own by means of a similar mechanism.

Suming up: the team of the National Corpus of Polish has to undertake following tasks as to secure representativeness of the corpus in question

1) set a text typology

2) set a typology of topics of print

3) reconstruct the structure of readership in Poland on the turn of the 20th century; the reconstruction shall be based first of all on researches already done.

4) "translate" the mentioned research done for different purposes into the text typology established for the purposes of the corpus.

REFERENCES

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1993). "Representativeness in corpus design". *Literary and Linguistic Computing* 8(4): 243-257.

Krâlik, J. & M. Śulc (2005). "The Representativeness of Czech Corpora". *International Journal of Corpus Linguistics* 10: 357-366.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K. & J. Woronczak (1990). *Słownik frekwencyjny polszczyzny współczesnej.* Krakow: Instytut Języka Polskiego PAN.

Sinclair, J. McH. & J. Ball: EAGLES Preliminary Recommendations on Text Typology. http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.htrnl.

Utka, A. (2004). *Statistical Identification of Text Functions.* Unpublished dissertation Kaunas Vytautas Magnus University.

Design and implementation of a CORpus di Italiano Scritto http://corpora.dslo.unibo.it/coris_engDesign.html.