The problem of balance and representativeness in the Electronic Corpus of 17th- and 18th-century Polish Texts

Ewa Rodek

Institute of Polish Language Polish Academy of Sciences

KorBa — Korba.edu.pl

The Electronic Corpus of 17th- and 18th-century Polish Texts, called also KorBa (Corpus of the Baroque era)

- Project developed by team from Institute of Polish Language, Polish Academy of Sciences (PAS) in cooperation with the Linguistic Engineering Group of the Institute of Computer Science (PAS). Manager: prof. dr hab. Włodzimierz Gruszczyński. Financing: National Program for the Development of Humanities
- **The first relatively large** corpus of old Polish texts (ca 22 mln of tokens)
- The only morphosyntactically annotated (including lemmatization) online corpus of pre-19th-century texts of such size in the Slavic world.
- **The main source** of developing Electronic Dictionary of 17th and 18th-century Polish (e-SXVII); second source is the archive of paper cards, which is available as a digitized images. Therefore, most of the documents selected for the 1st edition are texts included in the original canon of sources placed in the card index of the dictionary
- **1st edition** (2013–2018): 10,7 mln of tokens; 718 texts form 1601 to 1772
- **2nd edition** (2019–2023): increasing the volume and extending the time range by years 1773-1800. 10,8 mln of tokens; 1316 texts, 75% materials transcribed manually and 25% automatically in Transkribus (but the most difficult – manuscripts and texts with gothic fonts)

Theory

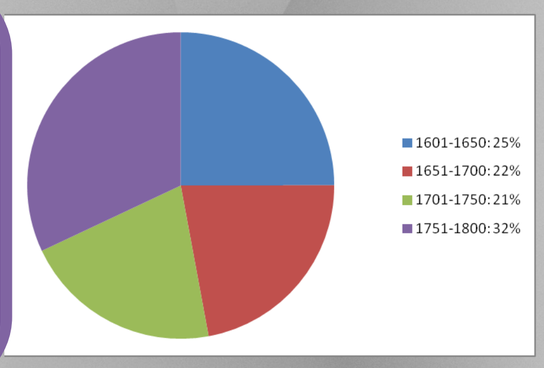
- **Representativeness** in corpus linguistics uses a properly selected text sample that can be generalized to the entire variety of the language (Leech 2001: 27).
- In historical corpora we can say rather about striving for representativeness and balancing rather than fully achieving these attributes (Adamiec 2015: 13; Krinková 2018: 13).
- This is due to **the literate nature** of the preserved historical documents (lack of representation for the oral variety of the language) and **the literary-centric** nature of the old literature. Works of belles-lettres have always been treated with greater attention, therefore they naturally constitute a significant part of the surviving texts. However, this category of texts cannot be allowed to be over-representative.
- In building diachronic corpora the big role plays **external, non-linguistic criteria** for the selection of texts
- **Balance** in diachronic corpus – the apparent paradox of the construction of large representative diachronic corpora: the diachronic corpus should be **heterogeneous** (i.e. contain texts by different authors, genres, styles, dialects), but it should be **homogeneous** in terms of individual chronology: periods should be comparable (preferably in the number of tokens also in the proportional representation of different text types) (Enrique-Arias 2012: 96)

Difficulties in gathering material

- **lack of material**: In the diachronic corpus, access to materials from the era is always limited. Many documents have not survived, some have been consumed by wars, fires or other vicissitudes.
- **re-editions after 1800**: Particularly difficult decisions concerned the inclusion of editions from the 19th century in the corpus, because the then editorial standards allowed publishers to significantly interfere with the original text, but these were the only surviving forms of texts that were interesting and worth introduction to the corpus. Taking into account contemporary editions of texts from the 17th and 18th century is associated with legal difficulties (restrictions resulting from copyright law) and different editorial principles of contemporary publishers, who often use more than one version of the original as the basis for editing.
- **small thematic diversity** of documents resulting from the specific historical and linguistic situation in that time: This made it very difficult to reconcile the representativeness of the material with the balance of the corpus.
- **placing manuscripts**: The problem is not only the state of preservation of the original and legibility of the writing, but also determining the date of creation and confirming the originality of the work.

Types of variation

Chronological: we divided material into four parts: 1601-1650, 1651-1700, 1701-1750 and 1751-1800. These timeframes are, of course, entirely artificial and serve only to order the material. The aforementioned volume criterion often had to be adjusted to account for other factors. Also we decided to divide the material on less artificial way on periods: 1601-1740 (Baroque – 60% of tokens) and 1741-1800 (Enlightment – 40% of tokens).

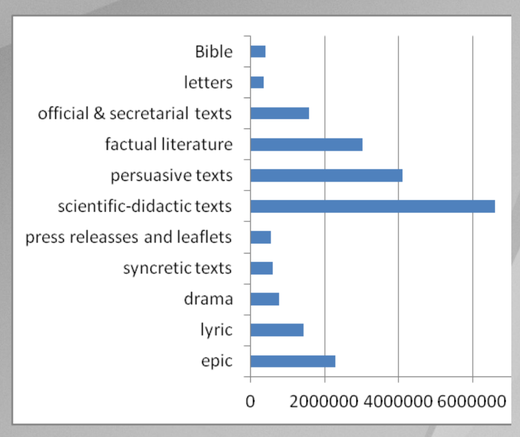


Geografical: The corpus includes texts from all regions where the Polish language was used, in accordance with most historical studies on this period. These are: Mazovia, Lesser Poland, Greater Poland, Ruthenian Lands, the Grand Duchy of Lithuania, Silesia, Livonia, Pomerania and Prussia.

In order to keep the corpus **balance**, very long texts have been included in fragments. Unfortunately, this had consequences for the material from modern chemistry and physics textbooks. They too had to be included in the excerpts, even though they were potentially the source of new words (neosemanticisms or scientific terms).

Genre: made up on several levels:

1. rhymed, non-rhymed and mixed texts,
2. literature and non-literary texts,
3. division into types and genres.



Polish literature background in the 17th & 18th C.

- **1st half of the 17th c.** saw the dynamic development of new intellectual trends of the Baroque era. The writings of this period are characterized by a variety of genres and topics, and freedom of expression. At that time, the literature of the high style experienced its heyday, as did the work of the so-called mockers – poets parodying high literature, playing with form, weaving elements of colloquial language into poetry.
- **The 17th century** was a time of constant wars. In the middle of the century, infectious diseases, climate cooling, droughts and floods prevailed throughout Europe, including Poland. All this led to the death of about 30% of Polish society, economic and, consequently, also cultural collapse of the whole country.
- The level of education, printing and reading deteriorated significantly. Publishing houses were taken over by religious congregations (mainly Jesuits and Piarists), which ran schools and needed school textbooks, but also materials for pervasive influence on society. Orders introduced strict censorship and published mainly religious works. As a result, the entire literature **from the mid-seventeenth century to the mid-eighteenth** century is monothematic – genres and topics related to the Catholic religion predominate (sermons, lives of saints, devotional, moralizing and liturgical works). Writers left their, often outstanding, works in manuscripts, there were no reissues of older works, which is why this period is called the age of manuscripts. Publishing production has decreased significantly: around 1650 ca 300-400 titles were published annually, between 1700-1710 – ca100-200 titles a year, between 1741-1750 ca 350 items a year (Imańska 2000).
- **It was not until the 1740s** that, there was a significant increase in publishing production and new Enlightment trends began to emerge in literature. Thematic and genre diversity has returned, especially in the field of applied literature. The second half of the 18th c. saw the rapid development of Enlightment literature, especially drama, novels, idylls and librettos, literature and the scientific press (chemistry, physics, mechanics, law, medicine).

Subject matter: we divided the documents into 48 thematic areas. Most popular were: religion, politics, history, alchemy, astrology, herbal medicine, phylosophy, but also metallurgy, aviation, nautica or culinary

References

- Adamiec D., 2015: Criteria for the selection of texts incorporated into "The electronic corpus of 17th- and 18thcentury Polish texts (up to 1772)", *Prace Filologiczne* 67, 11-20.
- Biber D., 1993: Representativeness in corpus design, *Literary and Linguistic Computing* 8/4, 243-257.
- Enrique Arias A., 2012: Dos problemas en el uso de corpus diacronicos del espanol: perspektiva y comparabilidad, *Scriptum Digital* 1, 82–106.
- Imańska I., 2000: *Druk jako wielofunkcyjny środek przekazu w czasach saskich*. Toruń: Publishing of Nicolaus Copernicus University.
- Krinková Z., 2018: Diachronní korpusová lingvistika a španělština: současný stav a problémy, *Časopis pro Moderní Filologii* 100, 1, 60–79.
- Leech G., 2001: Corpora. In: K. Malmkjaer (ed.), *The linguistics encyclopaedia*, 84-93. Routledge.