SIR Publishers
CENTER FOR INTERNATIONAL SCIENTIFIC PUBLICATIONS
JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN(Online): 2984-6722
SJIF Impact Factor |(2024): 6.752|
Volume-7, Issue-4, Published |20-04-2024|

ORAL CORPORA FOR BILINGUAL AND PLURILINGUAL CONTEXTS

https://doi.org/10.5281/zenodo.10972067

Sh. O'mrzoqov
JSPU, Jizzakh, Uzbekistan

SUMMARY

This article will revisit the discussion regarding small and large corpora, followed by an examination of bi- and plurilingual corpora. Initially, we will address the challenges associated with small bilingual oral corpora, including the comparability and generalizability of analysis results. Subsequently, we will present a specific example of a small bilingual corpus comprising linguistic data in German and English from various provinces. This corpus was utilized as a research database to investigate linguistic changes resulting from language contact.

Keywords

linguistic corpus analysis, bilingualism, multilingualism, comparable datasets.

Over recent decades, corpus linguistics has experienced a significant surge in significance. This surge can be attributed to advancements in IT, which have substantially enhanced our capabilities in data gathering, processing, and analysis. However, corpus linguistics still contends with diverse terminologies and methodologies. A corpus can broadly be construed as a compilation of language data organized according to specific criteria. Nevertheless, there remains a lack of consensus regarding the definition and boundaries of corpora in relation to other types of data collections, such as text compilations (including newspapers, letters, spoken language recordings, etc.), books, or web pages. Consequently, it's not always evident what attributes a text or linguistic data collection must possess to be classified as a "corpus." This issue is particularly pertinent concerning small corpora, as there's no clear delineation specifying the requisite length or quantity of texts needed to qualify as a "corpus."

In addition to the vague definition of the term "corpus," we also encounter ambiguity in the term "corpus linguistics." On one hand, corpus linguistics pertains to the creation of a corpus, specifically involving processes such as data collection, transcription, annotation, and standardization. Particularly, the latter aspect can be regarded as a distinct research area within the realm of automatic language processing (ALP). On the other hand, "corpus linguistics" can denote the utilization of corpora as empirical tools for research endeavors. Within this context, corpus

SIR Publishers
CENTER FOR INTERNATIONAL SCIENTIFIC PUBLICATIONS
JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN(Online): 2984-6722
SJIF Impact Factor |(2024): 6.752|
Volume-7, Issue-4, Published |20-04-2024|

linguistics integrates with various theoretical branches of linguistics to analyze working hypotheses or research questions using natural language data.

These inquiries underscore the significant role that corpus linguistics plays in contemporary linguistic research, revealing it to be a broad and expansive field of study. This paper aims to elucidate the theoretical and methodological challenges inherent in a specific type of corpus: small bilingual and plurilingual oral corpora.

Bilingual and plurilingual corpora hold paramount importance across various linguistic disciplines. They serve as invaluable resources in research on language contact and change, second or foreign language acquisition, linguistic variation, code-switching and mixed languages, migration linguistics, as well as psycholinguistic investigations of numerous languages. Nonetheless, the definitions of bilingual and plurilingual corpora are diverse, and the terminology often remains heterogeneous. This diversity inevitably complicates the efforts of researchers seeking to compile or locate a bilingual or multilingual corpus that meets their requirements. Consequently, this section is dedicated to presenting and defining bilingual and multilingual corpora, followed by an introduction to the specific type of corpus underpinning this contribution.

The scale is an inherent aspect of a bilingual or multilingual corpus. Backus has already highlighted various challenges associated with bilingual and multilingual corpora, which will be elaborated upon later in this section. Moving forward, we will begin by providing a terminological overview of the definitions of bilingualism and plurilingualism. Subsequently, it is imperative to precisely distinguish between different types of bilingual and multilingual corpora and compare their respective characteristics. In the subsequent section, we will delve into applying theoretical considerations to methodological challenges, specifically addressing the issues pertaining to the compilation and creation of bilingual and plurilingual corpora.

In broad terms, bilingualism typically involves proficiency in two languages, while plurilingualism denotes proficiency in more than two languages. Consequently, a corpus containing linguistic data in two languages is classified as a bilingual corpus, whereas one containing data in more than two languages is considered a multilingual corpus. An alternative perspective suggests categorizing a corpus as multilingual when it comprises more than one language, and treating it as monolingual when it contains only a single language.

However, the terms bilingualism and plurilingualism encounter additional terminological ambiguities. According to Oskar, a bilingual individual can easily switch between languages as required by the situation, implying comparable or similar proficiency in both (or multiple) languages. Yet, this is not always the case

SIR Publishers
CENTER FOR INTERNATIONAL SCIENTIFIC PUBLICATIONS
JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN(Online): 2984-6722
SJIF Impact Factor |(2024): 6.752|
Volume-7, Issue-4, Published |20-04-2024|

when referring to bilingualism or plurilingualism. Consequently, Lüdi identifies several factors influencing individual plurilingualism, with timing of language acquisition being perhaps the most significant aspect. Lüdi distinguishes between simultaneous acquisition, where a child learns two mother tongues from birth, and successive acquisition, involving the learning of a first mother tongue from birth followed by acquisition of a second mother tongue later on.

The discourse on terminology serves as a fundamental underpinning for the exploration of various categories of bilingual and plurilingual corpora. As previously noted, there exists a degree of ambiguity surrounding the definition of what constitutes a bilingual or plurilingual corpus. Researchers delineate three primary types of corpora encompassing multiple languages:

- Source texts accompanied by translations.
- Monolingual sub-corpora assembled under the same sampling framework.
- A fusion of both A and B.

On the flip side, a comparable corpus refers to a collection of linguistic data that shares similarities in terms of sample type and representativeness. This encompasses factors such as genre, text proportions, and publication periods. Parallel and comparable corpora serve distinct analytical purposes, such as translation or contrastive studies, and involve different approaches in their creation. In creating a comparable corpus, selecting an appropriate sample is crucial, whereas a parallel corpus may not require such adaptation since it consists of translations from a predetermined set of texts. A key challenge with small bilingual and multilingual corpora lies in their comparability, which affects the ability to generalize research findings. This issue stems not only from the corpus size but also from inadequate preparation and alignment of data, as well as the absence of universal transcription conventions. Moreover, small bilingual and multilingual corpora are typically assembled by individual researchers for specific projects, further complicating their comparability.

In summary, achieving consistency among small bilingual and multilingual corpora demands the availability of universal tools and resources for ensuring equivalence in data transcription and annotation. Moreover, it is essential to categorize various small bilingual and multilingual corpora based on distinct subjects and corpus types. Clearly, oral bilingual corpora with well-balanced bilingual speakers cannot be directly compared to parallel written corpora or comparable corpora in scientific jargon. Hence, careful consideration of corpus type and speaker characteristics is imperative prior to any comparison of bilingual and multilingual corpora. Organizing corpora by thematic relevance can prove immensely beneficial in this context, as exemplified by the corpus collection. The

project concentrates on small corpora of languages in contact and amalgamates 85 corpora showcasing "plurilingual exchanges between 30 languages and 180 speakers." This meticulously curated corpus collection facilitates the examination of contact phenomena such as borrowing and code-switching.

BIBLIOGRAPHY:

1. Gries, S. T. Quantitative corpus linguistics with R: A practical introduction. Routledge. 2020
2. O'Keeffe, A., McCarthy, M., & Carter, R. From Corpus to Classroom: Language Use and Language Teaching. Cambridge University Press. 2020
3. Sokolova, S. Yu. Corpus research of Russian grammar: problems of theory and methodology. Nizhny Novgorod: Nizhny Novgorod State University named after. N.I. Lobachevsky. 2021
4. Kavilova, Laura A Q U I L I N A. “THE CASE OF PREFIXATION IN UZBEK AND FRENCH NEOLOGISMS.” PEDAGOGICAL SCIENCES AND TEACHING METHODS (2023): n. pag. Print.
5. Kavilova, T. (2022). A MODERN TECHNOLOGIES IN TEACHING FOREIGN LANGUAGES DIGITAL GENERATION OF STUDENTS: MODERN TECHNOLOGIES IN TEACHING FOREIGN LANGUAGES DIGITAL GENERATION OF STUDENTS. Журнал иностранных языков и лингвистики, 4(4)