Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (1): 127-141, 2016 / EISSN:2215-2636

LINGUISTIC CORPORA OF UNDERSTUDIED LANGUAGES: DO THEY MAKE SENSE?

Corpus de lenguas poco estudiadas: ¿Tiene sentido?

*Igor Vinogradov**

ABSTRACT

A corpus of an understudied language usually has documentary-linguistic nature and comprises all text material available in a particular language. However, without resorting to text selection, it is impossible to obtain a representative and balanced sample of language use. Lack of these two characteristics makes a corpus almost useless for any kind of quantitative research. Nevertheless, corpora of understudied languages comply with a wide range of language documentation objectives. Furthermore, they can serve as evidence of the existence of word forms or grammatical features in texts that meet specific search criteria. If such corpora have well-elaborated linguistic annotation, they can complement grammatical descriptions and dictionaries, standing out against common text collections due to their digital format. They are especially suitable for typological research, when one has to deal with a huge amount of data in different and unrelated languages.
Key Words: corpus linguistics, understudied languages, language documentation, quantitative methods.

RESUMEN

Los corpora de lenguas poco estudiadas comúnmente surgen de las tareas de documentación lingüística y contienen todos los textos disponibles en una lengua particular. No obstante, sin seleccionar textos, no es posible obtener una muestra representativa ni equilibrada del uso de la lengua. Falta de estas dos características hace el corpus casi inútil en estudios cuantitativos. Sin embargo, los corpora de lenguas poco estudiadas cumplen con diferentes objetivos de documentación lingüística. Aparte, también sirven de evidencia de la existencia de formas de palabras o rasgos gramaticales en los textos que satisfacen criterios específicos de búsqueda. Si tienen anotación lingüística bien elaborada, pueden complementar descripciones gramaticales y diccionarios, distinguiéndose de las colecciones comunes de textos por su formato digital. Son particularmente útiles para estudios tipológicos, cuando uno tiene que tratar multitud de datos en diferentes lenguas.
Palabras clave: lingüística de corpus, lenguas poco estudiadas, documentación lingüística, métodos cuantitativos.

* Universidad Nacional Autónoma de México. Becario del Instituto de Investigaciones Antropológicas. México. Correo electrónico: happyjojik@yandex.ru
The author gratefully acknowledges support by the Program of Postdoctoral Fellowships at the National Autonomous University of Mexico.
Recepción: 27/01/16 Aceptación: 19/02/16.

1. Introduction

A number of small corpora of understudied or endangered languages from all over the world have appeared in the past decade (see Scannell 2007, Ostler 2008, Cox 2011 among many others). This paper presents a theoretical discussion of the application of such corpora in linguistic research.

McEnery and Ostler (2000: 403) claim that "if corpus linguistics is a useful approach in linguistics, then it should be applied to all languages". But complying with this imperative is not straightforward. The principles of building a corpus of a major national language and of a small understudied language are different. The main methodological problem with small corpora is the very limited selection of available text materials. This means that one of the basic tasks of corpus linguistics, namely "to make it possible to generalize from a corpus to a language as a whole or at least to a particular variety, register etc." (Gries 2009: 7), cannot be fulfilled.

There is no universally accepted conception of what a linguistic corpus is, nor is it obvious how to identify understudied languages compared to well-studied ones. The basic definitions adopted here are introduced in Section 1. Section 2 presents a brief overview of some examples of linguistic corpora of understudied languages from different genetic families and geographical areas (see also Ostler 2008). Section 3 describes other research instruments for comparison, covering corpora of major well-studied languages (3.1), language archives (3.2), and printed collections of annotated texts (3.3). Section 4 provides some ideas about possible research applications of the corpora of understudied languages. It is argued that text samples included in such corpora for objective reasons do not represent the variability of the language. Thus, quantitative methods of linguistic analysis based on the data from such corpora are not able to provide reliable results. However, the corpora of understudied languages are very useful in many other ways, as discussed in Section 4. Conclusions from this study are presented in Section 5.

1.1. Linguistic corpora

McEnery and Wilson (2001: 29) state that "in principle, any collection of more than one text can be called a corpus". Some authors adhere to this broad interpretation as referring to every text collection. The interpretation enables the definition of "corpus" to be expanded to encompass the entire Web (Kilgarriff and Grefenstette 2003) or, for instance, Web-based text collections of particular languages (Scannell 2007).

However, as McEnery and Wilson (2001: 29) rightly note, "in the context of modern linguistics" the term tends to be used in a more narrow sense. They consider four specific connotations of "corpus": sampling and representativeness, finite size, machine-readable form, and standard reference, i.e., wide availability to its potential users. By "sampling and representativeness" they refer to filling the corpus with "samples of a broad range of different authors and genres which, when taken together, may be considered to 'average out' and provide a reasonably accurate picture of the entire language population in which we are interested" (ibid.: 30). Thus, the narrow interpretation of a linguistic corpus can be phrased in the following way: "a finite-size body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" (ibid.: 32).

Other authors sometimes use other criteria to describe a "prototypical corpus". Gries and Berez (2015), for instance, enumerate four characteristics that "jointly define a prototypical corpus", which are slightly different from McEnery and Wilson's (2001) definition. Besides machine readability and representativeness, a "corpus is meant to be balanced, which means that the sizes of the subsamples (of speakers, registers, varieties) are proportional to the proportions of such speakers, registers, varieties, etc. in the population the corpus is meant to represent". Furthermore, a "corpus contains data from natural communicative settings, which means that at the time the language data in the corpus were produced, they were not produced

solely for the purpose of being entered into a corpus, and/or that the production of the language data was as untainted by the collection of those data as possible". Gries and Berez (2015) define linguistic corpus as "a category that contains exemplars that are prototypical by virtue of exhibiting several widely accepted characteristics, but that also contains many exemplars that are related to the prototype or, less directly, to other exemplars of the category by family resemblance links".

Consequently, there is a continuum of particular exemplars, ranging from the more prototypical, which satisfy all the characteristics, to the less prototypical. The British National Corpus can be considered prototypical: it is machine readable, of finite size, representative, balanced, widely available and contains data from natural communicative settings. It is important to note the possibility of deviations from the prototype. That is to say that if a specific text collection does not comply with one or more of the criteria, such as with collections in an understudied language, this does not automatically mean that it should not be treated as a corpus.

1.2. Understudied languages

There are no universally accepted criteria on how to delimit understudied languages from well-studied ones. For the purposes of this paper, understudied languages are equated with under-resourced ones, which may theoretically be inaccurate but in practice seems fair enough. The point is that there is (or was until recently) no continuous text production in these languages, which have no established literary tradition, although they may have a recent written tradition.

Szymanski (2011: 1) defines "resource-poor" languages as those that "lack any significant digital presence[1]". This study will not assume that materials should necessarily be digitized, because in principle every non-digital unit of text can technically be converted into a digital one, with more or less effort and expense. Szymanski (2011: 8) also notes that "resource-poor languages are not necessarily endangered, under-studied, or minority languages (although they may be)". Probably, under-resourced languages are still always understudied, because without significant resources it is hard to imagine a particular language producing many scientific works. The opposite seems also to be true, because available material almost always attracts scientific attention.

Maxwell and Hughes (2006: 30) use the term "lower-density" languages to refer to under-resourced languages. They draw a distinction between "high-density", "medium-density" and "lower-density" languages, attributing to the latter the absolute majority of the world's languages.

> For a few languages of the world (such as English, Chinese and Modern Standard Arabic, and a few Western European languages), resources are abundant; these are the high-density Languages. For a few more languages (other European languages, for the most part), resources are, if not exactly abundant, at least existent, and growing; these may be considered medium-density languages. Together, high-density and medium-density languages account for perhaps 20 or 30 languages, although of course the boundaries are arbitrary. For all other languages, resources are scarce. (Maxwell and Hughes 2006: 30)

2. Examining documentary corpora

The corpora of understudied and under-resourced languages usually have documentary-linguistic nature, since they are commonly "based on audio and video recordings that are transcribed, annotated, and described with metadata by either a single researcher working in the field or by a small team of researchers" (Gries and Berez 2015: 2). Such corpora do not meet some of requirements for "prototypical" linguistic corpora; see Section 1.1. A documentary corpus is always much smaller than a corpus of a major high-status language. Such corpora cannot be representative for a particular kind of speaker, register, nor language variety. And they neither can be balanced. Sometimes documentary corpora are not available to researchers who

did not participate in their creation. Sometimes they may contain language data that were produced especially for being included in a corpus, rather than being derived from natural communication. Thus, the only two mutual characteristics of a corpus of an understudied language and a prototypical linguistic corpus are the most basic ones: that their content is machine readable and their size is finite. This comparison is illustrated in Table 1.

TABLE 1.

Compliance of a documentary corpus with prototypical corpus characteristics

Characteristic	Compliance
Machine readability	+
Finite size	+
Representativeness	–
Balance	–
Data from natural communicative settings	–/+
Availability to researchers	–/+

In addition, an understudied language often lacks generally accepted standards of description fixed in a normative grammar and dictionary. Consequently, any kind of linguistic annotation (if we deal with an annotated corpus) depends on the subjective theoretical approach taken by a specific researcher. Although the presence of metadata does not appear in the list of basic features of a corpus, as will be argued in Section 4, the linguistic annotation (markup) is especially welcome for corpora of understudied languages.

The next section examines five corpora of understudied languages from Africa and Eurasia: firstly, the non-annotated corpora of Assamese and Ndebele (2.1), and then the annotated corpora of Ossetic, Bambara and Kalmyk (2.2). All these corpora are accessible freely via the Internet.

2.1. Examples of non-annotated corpora: Assamese and Ndebele

Assamese is an Indo-Iranian language spoken by almost 13 million people in India[2]. Written Assamese makes use of Bengali script. The Assamese corpus (https://cqpweb.lancs.ac.uk/ asm_v2, accessed 07-01-2016) was originally gathered by the Institute of Applied Language Sciences at Bhubaneshwar and then integrated in the scope of the EMILLE (Enabling Minority Language Engineering) project at Lancaster University and Sheffield University[3].

The Assamese corpus contains about three million tokens from 1,191 texts in total. All texts are divided into categories depending on the topic (e.g., business, education, mathematics), and users can exclude some of these categories by specifying a subcorpus. This corpus does not provide any kind of word-level linguistic annotation[4]. Therefore, the unique searchable items in the corpus are exact word forms or their parts. The developers of the interface do not provide a virtual keyboard, which could have been very useful taking into account the specific script.

Ndebele is a language belonging to the Bantu group of the Niger-Congo macro-family. It is spoken primarily in Zimbabwe by approximately 1.5 million speakers. The corpus of written and spoken Ndebele (http://www.edd.uio.no/allex/corpus/africanlang.html, accessed 03-01-2016) was developed within the ALLEX (African Languages Lexicon) Project. This corpus contains 691,268 tokens and is not annotated. A user has very few search options: one can use some regular expressions in the query and modify the extended context from 30 up to 1,000 symbols.

Hadebe (2002: 167) mentions that "the corpus consists of both oral and written texts, all transcribed and converted into machine-readable texts". The approximate percentage is 80/20% for written and oral parts, respectively. The oral part of the corpus violates the principle of "natural communicative settings" (see Section 1.1) because "most of the oral material

was collected by means of structured and unstructured interviews" (ibid.: 164), specially for inclusion in the corpus. For more information on the process of collecting and elaborating corpus material see the description of the Ndebele corpus in Hadebe (2002).

2.2. Examples of annotated corpora: Ossetic, Bambara and Kalmyk

The Ossetic language belongs to the Iranian branch of the Indo-Iranian subgroup of the Indo-European family. It is spoken by approximately 550,000 people in the Russian Federation and in Georgia.

The written corpus of the Ossetic language (http://corpus.ossetic-studies.org, accessed 02-01-2016) comprises more than 11 million tokens. It is a literary language corpus because it is basically formed of texts from literary journals as well as some works by Ossetic writers of the 20th century. The complete list of texts included in the corpus is provided on the corpus webpage. The texts belong to the Iron dialect of Ossetic, which is the basis for standard Ossetic.

This corpus is annotated; it includes grammatical information about tokens, as well as their translation. The annotation was made automatically and not disambiguated. The main merit of the Ossetic corpus is the powerful search engine[5], together with the user-friendly interface. One can search by lexeme, word form, translation, or by a particular set of grammatical features. One can include more than one token in the query and indicate distance between them. There are also some options to specify a subcorpus based on genre, period, authors and titles of documents, etc. Finally, a virtual keyboard is provided for non-standard symbols.

The authors of the corpus also provide some useful facilities to process the search results. It is possible to choose the output characters (Cyrillic/transliteration), the output layout (e.g., with or without morphemic annotation), the number of sentences in the expanded context, etc. The user can also sort the list of results by different parameters, including, for instance, the title/year of the document or preceding word form.

Bambara is a Mande language which belongs to the Niger-Congo macro-family. This language is not endangered (though it is still under-resourced), since it is spoken by more than 10 million people in Western Africa, generally in Mali.

The referential corpus of Bambara (http://cormand.huma-num.fr, accessed 03-01-2016) is annotated and contains both disambiguated and non-disambiguated subcorpora. As of October 2015, the total volume of the corpus amounts to almost three million tokens, while the disambiguated part is considerably smaller: 426,813 tokens. The corpus of Bambara is formed by texts from different sources that represent different genres and dialect zones. The interface of the website allows a user to specify a subcorpus in order to exclude some documents from the search.

The Bambara corpus allows users to search by lemmas, word forms, phrases, symbols, to specify part of speech and to set a specific context to the left, to the right or to the both sides. It is possible to visualize the results by plotting a frequency diagram.

Kalmyk is Mongolic language spoken in the southern part of the Russian Federation. According to Ethnologue (Lewis et al. 2015), there are 80,500 speakers of Kalmyk. The webpage of the Kalmyk corpus (http://kalmcorpora.ru, accessed 03-01-2016) reports that the situation of endangerment is even graver than it appears, since of these speakers no more than 5,000 are fluent.

As of May 2015, the Kalmyk corpus comprises the total of 8,691,671 words. The corpus is annotated morphologically and semantically, but not disambiguated. Users can use annotation in search queries. They also can limit the entire search to some particular genres, authors or text types (e.g., oral, folklore, poetic).

At this stage, the Kalmyk corpus does not provide much numerical or statistical information. The presentation of search results includes neither numbering nor the total

sentences found. The sections about statistics and frequencies on the corpus webpage are empty, although some graphic representations of statistical data about lemmas' and word forms' frequencies can be accessed via the link "Graphs".

Kukanova (2011) expresses an optimistic view of the possibility of obtaining a representative and balanced Kalmyk corpus in future. However, due to the general insufficiency and inadequacy of available text materials, it is difficult to share this opinion.

2.3. Overview

The corpora of understudied languages differ in numerous parameters. The composition of such corpora does not depend on the ideal theoretical conception of corpus structure, but more fundamentally on the quality, quantity and diversity of available texts in a particular language. This drastically affects the volume of the corpus and the coverage of different registers, genres and dialects. Corpus creators can solve the problem of coverage by intentionally provoking speakers to produce particular kinds of text that are lacking, as with the oral part of the Ndebele corpus (see Section 2.1). Nevertheless, one should be aware that in these cases the texts do not come from natural communication, and they may therefore be inappropriate for inclusion in a corpus.

Table 2 presents a short comparison of the five corpora examined in this Section.

TABLE 2.

Basic characteristics of the different corpora of understudied languages

Parameter	Ossetic	Bambara	Assamese	Ndebele	Kalmyk
1. Machine readable format	+	+	+	+	+
2. Finite size	+	+	+	+	+
3.Total volume (in million tokens)	~11	~ 3	~ 3	~ 0.85	~ 8.5
4. Morphologically annotated	+	+	–	–	+
5. (Partly) disambiguated	–	+	N/A[6]	N/A	–
6. Variety of search options	+	+	–	–/+	+
7. Limiting by a subcorpus	+	+	+	–	+
8. Facilities of result processing	+	+	–	–	+/–

The characteristics 4–8 in Table 2 do not deal with the basic features of a linguistic corpus discussed in Section 1.1. Annotation, disambiguation, variety of search options, possibility of creation of a subcorpus, special facilities to process the results – all these characteristics make a corpus more useful and more suitable for a wide range of research questions, but none of them can transform a simple raw text collection into a linguistic corpus.

All the examined corpora satisfy the criteria of machine readability and finite size (Table 2, parameters 1 and 2). But none of them satisfies the criteria of representativeness and balance. Furthermore, only some authors of documentary corpora discuss, in very restrained

and discreet fashion, the probability of the corpus in question being representative and balanced. The parameter of corpus volume (Table 2, parameter 3) does relate to these two criteria, but very indirectly; cf. "typically researchers focus on sample size as the most important consideration in achieving representativeness" (Biber 1993: 243). In fact, a huge range of sociolinguistic information should also be taken into account.

3. Comparable research tools

The documentary linguistic corpora examined in the previous section can be compared with some other research instruments, including corpora of major national languages, language archives and printed text collections. This section provides an overview of these. The differences between these tools and small corpora of under-resourced languages are addressed in Subsection 3.4.

3.1. Major national corpora

Xiao (2008: 383) states that "national corpora are normally general reference corpora which are supposed to represent the national language of a country". It is assumed that national corpora are usually highly developed and dispose of all conceivable tools and engines for successful research in different subdisciplines of linguistics. The search facilities provided in the corpus interface can therefore be ignored here.

The most influential example of a "large" linguistic corpus is the British National Corpus (BNC; http://www.natcorp.ox.ac.uk, accessed 04-01-2016). Aston and Burnard (1998: 28) state that "the BNC was designed to characterize the state of contemporary British English in its various social and generic uses". The starting point was the notion about an ideal language corpus without regard to the availability of text material. Particular material for inclusion in the corpus was selected later, based on decisions concerning corpus design, structure and predefined target proportions.

> The BNC project started with a careful planning stage where the design principles for the corpus were drawn up. These established a number of selection criteria which were then used for identifying suitable texts to be included in the corpus. (http://www.natcorp.ox.ac.uk/corpus/creating.xml, accessed 04-01-2016)

> It was hoped to maximize variety in the language styles represented, both so that the corpus could be regarded as a microcosm of current British English in its entirety, and so that different styles might be compared and contrasted. Each selection feature was divided into classes and target percentages were set for each class. Thus for the selection feature 'medium', five classes (books, periodicals, miscellaneous published, miscellaneous unpublished, and written-to-be spoken) were identified. Samples were then selected in the following proportions: 60 per cent from books, 30 per cent from periodicals, 10 per cent from the remaining three miscellaneous sources. Similarly, for the selection feature 'domain', 75 per cent of the samples were drawn from texts classed as 'informative', and 25 per cent from texts classed as 'imaginative'. (Aston and Burnard 1998: 29)

The 100-million-token British National Corpus complies with the criteria of being representative and balanced for a particular kind of speaker, register, variety, etc. The only significant point where the principle of balance has been intentionally violated is the ratio of the volume of the written corpus to the volume of the oral one.

> There is a broad consensus among the participants in the project and among corpus linguists that a general-purpose corpus of the English language would ideally contain a high proportion of spoken language in relation to written texts. However, it is significantly more expensive to record and transcribe natural speech than to acquire written text in computer-readable form. Consequently the spoken component of the BNC constitutes approximately 10 per cent (10 million words) of the total and the written component 90 per cent (90 million words). These were agreed to be realistic targets, given the constraints of time and budget, yet large enough to yield valuable empirical statistical data about spoken English. (http://www.natcorp.ox.ac.uk/docs/URG.xml, accessed 04-01-2016[7])

There is also a restriction for too-long texts, which are not be included entirely and truncated at least to 45,000 words.

The Russian National Corpus (http://ruscorpora.ru/en/index.html, accessed 04-01-2016) contains more than 300 million words. Like the British National Corpus, it can serve as an example of a representative and balanced corpus.

> A national corpus (…) is characterized by representative and well-balanced collections of texts. This means that such a corpus contains, if possible, all the types of written and oral texts present in the language (various genres of fiction, journalistic, academic, and business, as well as dialectal and sociolectal, texts). The proportion of text types in the corpus is based on their share in real-life usage at the time of composition. (http://ruscorpora.ru/en/corpora-intro.html, accessed 04-01-2016[8])

A more interesting example of a major national corpus is the Eastern Armenian National Corpus (EANC; http://www.eanc.net, accessed 04-01-2016). Armenian is the statutory national language of Armenia, a relatively small country in the Caucasus. Armenian belongs to the Indo-European family and is spoken by about six million people (Lewis et al. 2015). This corpus chooses its collection, so to speak, "semi-selectively".

> EANC is designed as a comprehensive corpus with the objective to include as many Standard Eastern Armenian texts as practicable. As of March 2009, EANC comprises about 110 million tokens. Overall, we have been guided by the goal of comprehensive representation – all literary, scientific and oral texts available to us have been indexed for search. The only exception to this are certain widely-available texts, such as electronic press and legal documents, whose presence has been limited for the sake of balance among different genres. (http://www.eanc.net/en/composition, accessed 04-01-2016)

The possible assortment of texts in Eastern Armenian, regardless of its official status, seems not to be large enough to allow corpus creators to reject some texts, trading the total volume for approximating to a more representative and balanced internal structure. Armenian is thus what McEnery and Ostler (2000) call a "smaller national language[9]".

3.2. Language archives

Another type of structured collection of linguistic data is the language archive. Language archives normally include different types of material, not only texts, and in that way they differ from language corpora. However, an archive usually includes all available materials, and the principles of sampling and representativeness are not relevant. These properties link archives with corpora of under-resourced languages, but not with major national corpora. Three archives are discussed below as examples.

One of the biggest language archives is the Archive of the Indigenous Languages of Latin America (AILLA; http://www.ailla.utexas.org, accessed 08-01-2016). It contains a wide range of linguistic data, from non-transcribed speech recordings and digitized researchers' field notes to morphologically analyzed texts, usually accompanied by morpheme-to-morpheme glossing and translation. The archive comprises data on more than 300 American indigenous languages, and most of the data are freely accessible. The AILLA webpage interface allows archives to be browsed by language, collection, country, and depositor's name.

Another example of a language archive is the Pangloss Collection (http://lacito.vjf.cnrs.fr/pangloss/index_en.htm, accessed 08-01-2016). The main goal of this archive is "to contribute to knowledge of endangered languages and cultures, by sharing annotated spoken texts of lesser-studied languages" (Michailovsky et al. 2014: 120).

> [The Pangloss Collection] contributes to the documentation and study of the world's languages by providing free access to documents of connected, spontaneous speech, mostly in endangered or under-resourced languages, recorded in their cultural context and transcribed in consultation with native speakers. The Collection is an Open Archive containing media files (recordings), text annotations, and metadata; it currently contains over

> 1,400 recordings in 70 languages, including more than 400 transcribed and annotated documents. The annotations consist of transcription, free translation in English, French and/or other languages, and, in many cases, word or morpheme glosses; they are time-aligned with the recordings, usually at the utterance level. A web interface makes these annotations accessible online in an interlinear display format, in synchrony with the sound, using any standard browser. The structure of the XML documents makes them accessible to searching and indexing, always preserving the links to the recordings. (Michailovsky et al. 2014: 119)

This archive is in fact not too far from a language corpus. It is fully digitized and open-accessed. It provides options for searching and indexing, and includes meta-information about recording and transcriptions.

Another language archive that also focuses on collecting audio- and video-ethnographic materials is the Ethnographic E-Research Online Presentation System (EOPAS; http://eopas.org, accessed 08-01-2016). EOPAS also provides interlinear linguistic analysis for its recordings[10], making them highly useful in many kinds of research. Currently this archive centers on indigenous languages from Australia and Oceania, and some from South America.

3.3. Printed collections of annotated texts

This is an old-fashioned method of representing linguistically analyzed texts in understudied languages. Such collections are usually quite small because of inevitable size limitations imposed by the printed format. For the same reason they are of course not digitized. Nevertheless, they may contain valuable information about language use and provide linguistic analysis of primary data. Very often, this information is not available elsewhere. The process of digitization (the technical details will not be addressed here) can enable the use of these sources in present-day computer-based search algorithms, thereby advancing the usability of printed text collections to the level of language archives or even corpora of understudied languages.

Printed collections of annotated texts are very widespread sources of linguistic data; there is therefore no need to cite particular collections here. Such collections can appear as separate books (for instance, Mayers 1958), as articles in specialized journals (Romero Méndez 2012) or as appendices to grammatical descriptions (Lacrampe 2014). The main object is to provide additional information on real-life language use that complements grammatical description and vocabulary.

3.4. Comparing different research tools

This Section presents a brief comparison of the four linguistic research tools examined above: small corpora of understudied languages, large national corpora, language archives and printed text collections. The basic parameter is the policy carried out regarding the selection of materials to be included in the research device. The developers of small corpora of understudied languages cannot afford the luxury of rejecting available texts, because such languages are normally under-resourced. The same is true for language archives that include material of every kind (sometimes not only textual) in order to cover the entire language use. The developers of large corpora, on the contrary, normally have an almost unlimited selection of available texts, so they have no problem leaving some of them beyond the scope of the corpus in order to maintain its representativeness and balance. Interestingly, authors of printed text materials usually have the same luxury of being selective, due to reasons of space.

Table 3 shows a comparison by this and some other parameters.

In fact, every group of instruments mentioned in Table 3 is very diverse. For example, small documentary corpora can be annotated or not, can have a more developed search engine or a less developed one, and can of course be bigger or smaller in size. Comparing the groups in Table 3, we can see that small documentary corpora are placed somewhere between major national corpora

at one end and language archives at the other, according to their basic characteristics. They share some features with the former and others with the latter. Ostler (2008), for example, includes language archives such as AILLA (Section 3.2) or OLAC (Open Language Archives Community; http://www.language-archives.org, accessed 08-01-2016) among corpora of less studied languages. A terminological trap whereby both small and large corpora are traditionally called "corpora", and archives are called "archives", should not confuse the matter.

TABLE 3.

Basic characteristics of different research instruments

Parameter	Major corpora	Documentary corpora	Language archives	Printed text collections
1.Selectivity of material	+	+	–	+/–
2.Machine readable format	+	+	+	–
3.Volume	Big	Small	Big/small	Very small
4.Morphological annotation	+	+/–	–/+	+/–
5.Search facilities	+	+/–	–/+	–

4. Discussion

It has been argued above that the major complaint against corpora of understudied languages is that they cannot be representative or well-balanced, unlike prototypical corpora of widespread national languages. The difference between a corpus of an understudied language and a language archive, incidentally, consists mostly in the consistency of data and the manner of presentation; in other words, in annotation, metadata, search facilities, etc.

The use of corpora of understudied languages undoubtedly makes sense. While not being as simple as language archives, they perform all the same duties of language documentation, i.e., they provide "a comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann 1998: 166) or, in other words, the "creation, annotation, preservation, and dissemination of transparent records of a language" (Woodbury 2011: 159[11]). A corpus of an understudied language is more than a language archive because it usually provides metadata on the included texts, some kind of linguistic annotation and useful search facilities. On the other hand, according to the narrow understanding of a linguistic corpus, such corpora are not even corpora because they do not comply with criteria of sampling, representativeness and balance. How, then, can they be used?

The primary objective of a linguistic corpus is "to help linguists find and explore sentences (occurrences) in texts [in a particular language] that meet specific search criteria" (http://www.eanc.net/en/objective, accessed 06-01-2016). This goal does not directly depend on the representativeness of the corpus,

but rather on its size and the perfection of the search mechanism. Consequently, it is a realistic objective for a corpus of an under-resourced language.

The next step is to make quantitative generalizations about corpus findings. Here we encounter a problem. As Heylen (2005: 261) rightly notes, any result of a quantitative corpus-based study is "strictly speaking only valid for observations that instantiate a similar type of language use as the one that was represented in the corpus". In other words, any kind of statistical information derived from a corpus which is not representative for a particular register, location, or time period does not make sense if applied to that register, location, or time period, or to the whole language. For example, the fact that 30% of word forms in the corpus of language X have the feature *n* by itself does not actually tell us anything about *X* until we carefully analyze the metalinguistic characteristics of the texts which form the corpus. Biber and Conrad (2001: 332) note that "although corpora are valuable for providing natural examples of words or grammatical features in context, corpus linguistics offers a unique perspective because of its use of quantitative analyses, which allow researchers to investigate patterns of language use that are otherwise impossible to ascertain". This is true only for representative and well-balanced corpora[12]. The corpora of understudied languages do not provide this perspective of quantitative analysis.

In fact, the situation is even more tricky. Regardless of corpus representativeness, it is still possible to apply quantitative methods when there are countable data of any sort in the corpus. But the results of such research will very probably be unreliable. Sometimes it can be quite difficult to resist the temptation to resort to quantitative methods regardless of the inappropriateness of the initial data. The corpora of under-resourced languages offer no possibility "to quantitatively test hypotheses about syntactic and semantic tendencies in language production", which some authors consider crucial in order to overcome "a serious methodological weakness affecting much research in syntax and semantics within the field of linguistics" (Gibson and Fedorenko 2010: 233).

The documentary corpora of under-resourced languages seem most suited to providing natural examples of language use[13]. The usability of such corpora highly depends on the coverage of linguistic annotation and the quality of the search mechanism. The more facilities are provided, the easier it is to find the word form or morpheme being looked for. This kind of application is especially welcome for typological research, when a linguist has to deal with a huge amount of data in different and often unfamiliar languages. Due to the digital format, a corpus is much more useful for this purpose than a printed collection of annotated texts. In this sense, a documentary corpus is an improved substitute for a simple collection of texts, which together with grammar and dictionary makes up a language description "triad" (see, for instance, Tsunoda 2006: 29).

5. Conclusions

An under-resourced language corpus is commonly documentary in nature and comprises all text material available in a particular language. Without resorting to text selection, it is impossible to obtain a representative and balanced sample of language use. Lack of these two characteristics makes a corpus almost useless for any kind of quantitative research. Nevertheless, it can still perform the primary function of a corpus, i.e., to serve as evidence of the existence of sentences or word forms in real texts that meet specific search criteria indicated by the researcher.

If such a corpus has well-elaborated linguistic annotation, it can be very useful in various kinds of research that do not presume quantitative methods. For example, such corpora are especially suitable for typological research, when one has to deal with a huge amount of data in different and unrelated languages. They complement grammatical descriptions and dictionaries, standing out against common printed (glossed) text collections due to their machine readable format and automatic search

facilities. Furthermore, the corpora of under-resourced languages should be considered as more powerful kinds of language archive. In this sense, they also comply with a wide range of language documentation objectives.

Notas

1. King (2015) calls such languages "low-resource", referring to the lack of available resources.

2. Here and below, basic information about languages is cited by Glottolog (Hammarström et al. 2015) and Ethnologue (Lewis et al. 2015).

3. See Baker et al. (2002) for more information on the EMILLE project.

4. However, some other corpora from the EMILLE project do include an annotated component; for instance, the Urdu texts are part-of-speech tagged (Baker et al. 2004).

5. This is the same search engine that was adapted from the Eastern Armenian National Corpus, see Section 3.1. For more technical details see Arkhangelskiy et al. (2012).

6. The parameter of disambiguation is applied only to morphologically annotated corpora.

7. Cf. also Leech (1992: 4): "there is an enormous imbalance between the amount of written and spoken corpus data available: something which is reflected in the composition of the BNC, of which only 10 million words at the most are likely to be of speech".

8. See also Sharoff (2006) for more details on the design of the Russian National Corpus.

9. McEnery and Ostler (2000: 407) estimate the population of speakers of a "small national language" to be under one million people, which is not true for Armenian.

10. For more technical details on the EOPAS system see Schroeter and Thieberger (2006).

11. Cf. also Cox (2011: 240): "corpus linguistics intersects with language documentation (…) inasmuch as it deals with the construction and analysis of consistent, reusable collections of linguistic data".

12. However, note the pessimistic view on representativeness even for English corpora in Manning and Schütze (2000: 21): "in general the goal of using a truly 'representative' sample of all of English usage is something of a chimera, and the corpus will reflect the materials from which it was constructed".

13. This fact, among other things, enabled Mosel (2014) to extend the range of possible applications of a documentary corpus to include the production of grammatical descriptions of previously undescribed languages.

References

Arkhangeskiy, Timofey, Oleg Belyaev and Arseniy Vydrin. (2012). "The Creation of Large-Scale Annotated Corpora of Minority Languages Using UniParser and the EANC Platform". In: *Proceedings of the 24th International Conference on Computational Linguistics* (December 2012, Mumbai, India): 83-92.

Aston, Guy and Lou Burnard. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Austin, Peter K. and Julia Sallabank (eds.). (2011). *Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

Baker, Paul et al. (2002). "EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation". In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (May 2002, Las Palmas, Spain).

Baker, Paul *et al*. (2004). "Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development". In: *Literary and Linguistic Computing,* 19 (4): 509-524.

Biber, Douglas. (1993). "Representativeness in Corpus Design". In: *Literary and Linguistic Computing,* 8 (4): 243-257.

Biber, Douglas and Susan Conrad. (2001). "Quantitative Corpus-Based Research: Much More Than Bean Counting". In: *TESOL Quarterly,* 35 (2): 331-336.

Cox, Christopher. (2011). "Corpus Linguistics and Language Documentation: Challenges for Collaboration". In: Newman, Baayen and Rice (eds.): 239-264.

Gibson, Edward and Evelina Fedorenko. (2010). "Weak Quantitative Standards in Linguistics Research". In: *Trends in Cognitive Sciences,* 14: 233-234.

Gries, Stefan Th. (2009). "What is Corpus Linguistics?" In: *Language and Linguistics Compass,* 3: 1-17.

Gries, Stefan Th. and Andrea L. Berez. (2015). "Linguistic Annotation in/for Corpus Linguistics". In: Ide and Pustejovsky (eds.): [in print].

Hadebe, Samukele. (2002). "The Ndebele Language Corpus: A Review of Some Factors Influencing the Content of the Corpus". In: *Lexikos,* 12: 159-170.

Hammarström, Harald *et al*. (2015). Glottolog 2.6. http://glottolog.org. Consulted: 03-01-2016.

Heylen, Kris. (2005). "A Quantitative Corpus Study of German Word Order Variation". In: Kepser and Reis (eds.): 241-263.

Himmelmann, Nikolaus P. (1998). "Documentary and Descriptive Linguistics". In: *Linguistics,* 36: 161-195.

Ide, Nancy and James Pustejovsky (eds.). (2015). *Handbook of Linguistic Annotation*. Berlin – New York: Springer.

Kepser, Stephan and Marga Reis (eds.). (2005). *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Berlin: Mouton de Gruyter.

Kilgarriff, Adam and Gregory Grefenstette. (2003). "Introduction to the Special Issue on the Web as Corpus". In: *Computational Linguistics* 29 (3): 333-347.

King, Benjamin Philip. (2015). Practical Natural Language Processing for Low-Resource Languages. Doctoral dissertation. University of Michigan.

Kukanova, Viktoria V. (2011). "General Structure and Perspectives of Application of National Corpus of Kalmyk Language in the Light of the Representativeness Problem". [in Russian] In: *Proceedings of the XL International philological conference* (March 2011, Saint-Petersburg, Russia): 125-137.

Lacrampe, Sébastien. (2014). Lelepa: Topics in the Grammar of a Vanuatu Language. Doctoral dissertation. Australian National University.

Leech, Geoffrey Neil. (1992). "100 Million Words of English: the British National Corpus (BNC)". In: *Language Research* 28 (1): 1-13.

Lewis, M. Paul, Gary F. Simons and Charles D. Fennig (eds.). (2015). Ethnologue: Languages of the World. 18th Edition. http://www.ethnologue.com Consulted: 03-01-2016.

Lüdeling, Anke and Merja Kytö (eds.). (2008). *Corpus Linguistics: An International Handbook*. Volume 1. Berlin – New York: Walter de Gruyter.

Manning, Christopher D. and Hinrich Schütze. (2000). *Foundations of Statistical Natural Language Processing*. 2nd Edition. Cambridge: MIT Press.

Maxwell, Mike and Baden Hughes. (2006). "Frontiers in Linguistic Annotation for Lower-Density Languages". In: *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora* (July 2006, Sydney, Australia): 29-37.

Mayers, Marvin. (1958). *Pocomchi texts with grammatical notes*. Norman: Summer Institute of Linguistics of the University of Oklahoma.

McEnery, Tony and Nick Ostler. (2000). "A New Agenda for Corpus Linguistics – Working with All of the World's Languages". In: *Literary and Linguistic Computing* 15 (4): 403-418.

McEnery, Tony and Andrew Wilson. (2001). *Corpus Linguistics: An Introduction*. *2nd* Edition. Edinburgh: Edinburgh University Press.

Michailovsky, Boyd *et al.* (2014). "Documenting and Researching Endangered Languages: The Pangloss Collection". In: *Language Documentation & Conservation* 8: 119-135.

Mosel, Ulrike. (2014). "Corpus Linguistic and Documentary Approaches in Writing a Grammar of a Previously Undescribed Language". In: *Language Documentation & Conservation* 8: 135-157.

Newman, John, R. Harald Baayen and Sally Rice (eds.). (2011). *Corpus-Based Studies in Language Use, Language Learning, and Language Documentation*. Amsterdam: Rodopi.

Ostler, Nicholas. (2008). "Corpora of less studied languages". In: Lüdeling and Kytö (eds.): 457-483.

Romero Méndez, Rodrigo. (2012). "Ja' Apokäät o la narración de la curandera". In: *Tlalocan* XVIII: 79-123.

Scannell, Kevin P. (2007). "The Crúbadán Project: Corpus Building for Under-Resourced Languages". In: *Cahiers du Cental* 5 (1): 1-10.

Schroeter, Ronald and Nicholas Thieberger. (2006). "EOPAS, the EthnoER Online Representation of Interlinear Text". In: *Proceedings of the Conference "Sustainable Data from Digital Fieldwork*" (December 2006, Sydney, Australia): 99-124.

Sharoff, Serge. (2006). "Methods and Tools for Development of the Russian Reference Corpus". In: *Language and Computers* 56 (1): 167-180.

Szymanski, Terrence D. (2011). Morphological Inference from Bitext for Resource-Poor Languages. Doctoral dissertation. University of Michigan.

Tsunoda, Tasaku. (2006). *Language Endangerment and Language Revitalization: An Introduction*. Berlin – New York: Mouton de Gruyter.

Xiao, Richard. (2008). "Well-known and influential corpora". In: Lüdeling and Kytö (eds.): 383-457.

Woodbury, Anthony C. (2011). "Language Documentation". In: Austin and Sallabank (eds.): 159-186.