UDC 81'32 https://www.doi.org/10.33910/2687-0215-2020-2-2-147-159

A corpus-based approach in archaeolinguistics

I. Afanasev[✉1]

[1] Saint Petersburg State University, 7/9 Universitetskaya Emb., Saint-Petersburg 199034, Russia

Abstract. The article focuses on archaeolinguistics as a separate field of knowledge and outlines the features that distinguish it from other disciplines in comparative studies. It analyses the existing text collections and shows how they may find application in a corpus-based research in ancient languages. It also discusses approaches to creating new corpora of texts. The study focuses on Old Church Slavonic and Ancient Greek, in particular, it analyses the existing corpora in these languages, e. g., Corpus Cyrillo-Methodianum Helsingiense. Most of the corpora under study are not tagged. Some of them change the original writing system (from Glagolitic to Latin, using, for instance, ASCII), while the others have a restricted access. Some of the corpora are no longer available at all or available as part of local databases only. Thus, corpus-based resources in ancient languages in question are obviously insufficient. To facilitate more effective research, the easiest possible solution is to develop new corpora by using platforms specializing in linguistic analysis (e. g., CDLI or Lingvodoc) or systems that support DIY corpora. However, such platforms are often paywalled, may have limited functionality, or lack comprehensive user guides. With all the above in mind, there seems to be no ready solution for archaeolinguists who want to use a corpus-based approach in their study. They either have to make a considerable effort to modify an existing system for their purposes, or to build one of their own. In conclusion, the article proposes one of the possible ways to address these issues.

Keywords: archaeolinguistics, corpus-based approach, review, Old Church Slavonic, Ancient Greek, corpus linguistics, ancient languages, extinct languages.

Аннотация. В статье дается определение археолингвистики как науки и ее особенности, отличающие ее от других дисциплин сравнительно-исторического языкознания. Главный предмет рассмотрения — корпусные методы в изучении древних языков, а также главный материал, на котором возможно применить данные методы, — существующие текстовые коллекции. Также изучается возможность создания новых корпусов. Языки, на материале которых исследование проводится, — старославянский и древнегреческий. Существующие корпуса данных языков, такие как PROIEL, коллекция текстов университета Франкфурта, или Хельсинкский Кирилло-Мефодиевский корпус, подвергаются анализу. Большая часть данных корпусов представляет собой неразмеченные (не содержащие, к примеру, информацию о части речи конкретных токенов) электронные коллекции. В некоторых из них изменена кодировка. Некоторые из этих корпусов прекратили свое существование или же были перемещены из открытого доступа в локальные базы данных. Наиболее простое решение проблемы недостатка ресурсов — самостоятельное создание таких корпусов, которые бы подходили для исследования, при помощи платформ, специализирующихся на лингвистическом анализе, таких как CDLI или Lingvodoc. Однако у таких ресурсов может не хватать документации, набора функций, способности адаптироваться к нуждам исследователей текстов, написанных с использованием иных систем письменности, или же возможности использования в различных системных окружениях. Следующим шагом стало рассмотрение систем, спроектированных с целью помощи исследователям в создании непосредственно корпусов. Они могут располагаться за пэйволлом, обладать недостаточно гибкими функциями или, как и в предыдущем случае, функционировать неполно в различных окружениях, чтобы удовлетворять нужды исследователей, работающих в различных операционных системах. Все эти обстоятельства означают, что готовых решений для исследований в области археолингвистики, опирающихся на корпусный метод, не существует. Археолингвистам потребуется или тщательно адаптировать уже существующие решения, или создавать новые. В заключении предлагается один из возможных способов решения данной проблемы.

Ключевые слова: археолингвистика, корпусные методы, обзор, старославянский язык, древнегреческий язык, корпусная лингвистика, древние языки, мертвые языки.

Introduction

Archaeolinguistics is a subfield of linguistics that attempts to reconstruct the language system without having a possibility to get data directly from native speakers (including information on how not to use some language entities). What archaeolinguistics has available for analysis are incomplete evidences from the language systems long gone: manuscripts and texts. Their linguistic study is paramount because they stand out in their genetic unities (branches, groups, even families) as the first written evidence of their existence. Thus, they provide researchers with insights into the earliest stages of a language lifecycle which facilitates a more effective analysis of differences between their contemporary relatives. Withal, data, acquired in this sort of research, may be helpful in the studies of modern languages *per se* (Eckhoff 2018).

The focus of archaeolinguistic research lies at the intersection with other linguistic disciplines, however, we have to mention some key differences. Archaeolinguistics and linguistics of particular extinct languages, such as, e. g., Old Church Slavonic, develops its own general research methodology in the absence of data obtained from native speakers and in view of the abundance of negative material. Palaeolinguistics, for example, tends to reconstruct the systems preceding the languages, not paying a lot of attention to the languages *per se*. In contrast with glottochronology and evolutionary linguistics, archaeolinguistics does not attempt to date particular issues or explore the language origins. Finally, as compared to historical linguistics, archaeolinguistics seems to be more interested in how a written language emerged than in its genetic background.

Corpora provide completely new possibilities for linguistic research (Zakharov 2015, 11). Today, most researchers in archaeolinguistics resort to DIY corpora of ancient texts and see numerous advantages in the use of annotated resources and enhanced search engines (Mitrenina 2014, 44). This tendency touches the languages that, having changed a bit, still function (Alrabiah et al. 2014), as well as the ones that died out completely (Molina, Molin 2016). The article aims to apply the outlined trends to the study of two languages: Old Church Slavonic and Ancient Greek.

Old Church Slavonic is the first written Slavic language ever existed. It is a written-only language, i. e., a written text was the only form of its existence. Hence, the development of a corpus of texts in Old Church Slavonic seems only natural and there have been several attempts to do that over the last 30 years. Here we can mention such corpora as (TITUS), (CCMH), (Obshtezhitie), (PROIEL), (USC ODC), and (RRuDi). However, neither of them was a success. As a result, the majority of studies in Old Church Slavonic are mostly cultural (cf., for instance, Vendina 2002). Along with that, the linguistic study of Old Church Slavonic is high on the research agenda for many disciplines as the language of Orthodoxy — one of the foundations of Russian culture. Here, no research is effective without a comprehensive corpus.

Old Church Slavonic was mostly based on the Ancient Greek originals, being, *en mass*, the language of translation (excluding a small number of texts). This necessitates the study of Ancient Greek as well, which, in fact, is one of the cornerstones of European culture.

Today, the biggest challenge for linguistic research in ancient languages is the insufficient number of ancient language corpora, namely, *tagged corpora* (Dandapat et al. 2004, 170).

If a tagged corpus was once compiled, it is either incomplete, unavailable, or deleted from the Internet. This is true not only about Ancient Greek, or Old Church Slavonic, it is true about archaeolinguistics in general (Sokolov 2019).

The article is organised as follows. Section 2 provides an overview of the existing corpora of Old Church Slavonic (2.1) and Ancient Greek (2.2). It also outlines their disadvantages. Section 3 discusses corpus building software aimed to enhance the effectiveness of research in Old Church Slavonic. Section 4 concludes the paper and highlights the ways to resolve the existing issues.

Corpora: A lifesaver for researchers in ancient languages

Old Church Slavonic

Several recent decades have witnessed a range of attempts to build Old Church Slavonic corpora. Here, we can mention such corpora as the University of South California corpus, the Regensburg corpus, the text collection of Goethe University, Corpus Cyrillo-Methodianum Helsingiense, PROIEL and part of the Universal Dependencies project based on it as well as the text collection of the Obshtezhitie project. All of them can benefit research in ancient languages, however, neither of them is perfect or sufficient.

The Old Church Slavonic corpus of the University of South California (USC OSC) has always provoked some sort of discussion, mainly, because of its highly limited accessibility. In fact, the corpus was virtually non-existent for much of the research — it took the author of this article over a year to get an approval of the request for its use. Now, an attempt to go to the query page returns Page not Found (error 404) from the University website — a machine-generated message that a server is unable to find data according to the query. Thus, the Old Church Slavonic corpus of the University of South California seems to no longer exist.

When it was still available in the restricted access mode, it was described as a comprehensive collection of texts, each PoS tagged. However, it was impossible to check if it was really so, hence, the corpus became totally irrelevant for all the further studies. The experience with the Regensburg corpus was similar (RRuDi).

Another collection of texts in Old Church Slavonic (Manuscript) offers a range of unique features, for example, full PoS tagging or lemmatization and the representation of texts in different writing systems. However, the access to the collection is granted only upon a request. This may potentially pose a threat to the preservation of this collection (and, consequently, possible reproduction of research results). Moreover, the collection does not include all the currently available texts in Old Church Slavonic, so it is neither balanced nor representative as the definition of a text corpus requires it to be.

The collection of Old Church Slavonic texts on (TITUS) and (CCMH) websites was edited by one and the same team and it features the links referring to the same page. For this reason, the overview below will discuss (TITUS), (CCMH) and the same texts collected in other sources, for example, (Obshtezhitie).

One of the obvious advantages of the collection is that it is openly available. The downside is that the list of sources is incomplete: both (TITUS) and (CCMH) lack some truly important ones. For example, (TITUS) does not include Codex Assemanius, while (CCMH) does not feature Kiev Leaflets. Despite this, the collection is still the biggest and instrumental in Old Church Slavonic studies.

Some sources on the website (Obshtezhitie) are partially morphologically tagged. Some texts in (TITUS) are searchable. Both (TITUS) and (CCMH) provide access to the original text.

None of the above sources provides a full PoS tagging even with unresolved ambiguity. In fact, all the work has been done manually by volunteers. The most complete source lacks PoS tagging, and there is no evidence it will be available any time soon. There are no automatic tagging tools developed for Old Church Slavonic. So, it is barely possible to extend the existing collections or create new ones without investing a disproportionate amount of effort.

To sum up, no collection from this group is an Old Church Slavonic corpus that matches the necessary criteria, such as completeness and representativeness.

The PROIEL corpus is highly valuable for the students of Old Church Slavonic. It is a parallel corpus of the New Testament with the PoS tagged Codex Marianus. The corpus makes it possible to compare the original Greek texts with the Old Church Slavonic translations. This makes the corpus exceptionally helpful as it may provide insights into the decisions made by the translators who actually developed the language.

The corpus was incorporated in the Universal Dependencies 2.6 collection (Zeman et al. 2020). Using its data, it is possible to make an attempt to create a machine learning tool to provide PoS tagging of the other texts with a relatively high precision. The format and the completeness of the collection data allow using it in existing models provided via open source licenses.

All things considered, though, it is the New Testament corpus built to analyse different interpretations of one particular text, and not the features of a language or even languages the New Testament was written in. Thus, its role as the source of Old Church Slavonic material is largely auxiliary. Besides, it is incomplete and unrepresentative, because some Old Church Slavonic texts are not part of the New Testament. Some of them are found in (TITUS) and (CCMH).

Ancient Greek

The Ancient Greek corpus PerseusDL (Perseus) is a text collection tagged up to the syntactic level (implying that, for instance, PoS tagging has already been done as well (Hasan et al. 2007, 1)). However, this corpus is now available only as a GitHub repository, which makes it hard to get for a researcher with no relevant technical expertise. It does not have a desktop or a web application either. Perseus is an XML file collection with tagging in its raw format. Turning it in a full-fledged corpus requires the effort of software engineers with relevant expertise. By way of a reminder, there is the New Testament corpus of Modern Greek (SBLGNT) with additional morphological data (Tauber 2017), Ancient Greek subcorpus of (PROIEL), and Duke Databank of Documentary Papyri (DDBDP). The first and the third represent different stages of the language development lifecycle, while the second comprises the New Testament texts only. Corpus (PROIEL) is thoroughly tagged, yet it lacks amount of material, being just the intersection for both Ancient Greek and Old Church Slavonic, incomplete for each of the languages. PerseusDL, on the contrary, is primarily a corpus of Ancient Greek with no biblical texts included.

Besides, Ancient Greek corpora are also represented by *Thesaurus Lingua Graecae* (Berkowitz, Johnson 1990). Its pros (the number of texts and lexical units, lemmatization), and cons (lacking critical apparatus, the texts comprising the corpora are not given in their optimal version), are thoroughly described by the Greek language researchers, who share the view that the cons do not allow to “see TLG as a full-fledged philological corpus” (Arkhangelsky, Kisilier 2018). In addition, access to TLG is based on a “subscription fee” (Arkhangelsky, Kisilier 2018), that makes it even less available to the researchers.

Lacking corpora of extinct languages: The reasons behind

In view of the above, a question begs itself. If the issue has been around for some time, why have not there been attempts to resolve it?

One can argue that one of the key obstacles is the manual deciphering of manuscripts: possibilities of OCR now are quite limited for this particular task. This is what hinders research and the transformation of a text collection into a corpus. However, this is not the biggest issue of concern for Old Church Slavonic as a certain number of long and important texts is, in fact, well represented in a machine-readable form. By and large, the issue seems to be quite common and has to be addressed separately for each particular language.

Probably, the key matter of contention is a lack of easy-to-use software to facilitate effective tagging of a big enough corpus, or, at least, its automated placement into a database and desktop/ web application.

The biggest issue is the tools for computer-aided tagging. No single existing Old Church Slavonic corpus has a software tool to effectively perform its tagging, even if we take pay-walled software or tools with limited access. Even if the corpus itself is tagged, it is done manually, thus, each following text takes more person-hours. The development of specialized software is also a challenging task because it requires the contribution of numerous experts from the specific branch of historical linguistics and the usage of advanced machine learning techniques.

A possible solution here is open source corpus-building software (Kopotev 2014, 110). Even a small research team may ensure a full-fledged adaptation of the programme for a particular language.

A brief overview of corpus-building open source software

This part focuses on open source software that meets the needs of archaeolinguistics. Firstly, it should allow to update the original product (the programme that builds a corpus based on a set of languages should have user interface for the addition of new ones). Secondly, it should use an accessible source code that enables the user to adapt the programme for individual needs. These two criteria are the decisive parameters in the overview that follows. The overview explores such projects as CDLI and Lingvodoc databases, Sketch Engine and Tsakorpus platforms as well as AntConc and #LancsBox toolsets.

CDLI

CDLI (The Cuneiform Digital Library Initiative) is a twenty year strong international project that curates cuneiform text artifacts in the interest of open access research (CDLI Core Update).

The project, as the name implies, is a huge digital library, created by the scholars from the USA, Canada, Germany and the UK (CDLI Core Update) (MTAAC). The website (CDLI) offers a substantive database of cuneiform documents as well as publications based on this material. The search options include such parameters as publication metadata, information about the collection, transliteration of a particular word, an ID in CDLI database. Each cuneiform document is annotated according to these parameters. The pages with documents feature photos so that a researcher may compare the photo with the transliterated version to avoid possible errors.

The project code is open source and is accessible via GitHub (CDLI Repository). Inside, there are forty-one repositories, with five of them specifically highlighted by the developers, i. e., data (most of the data, accessible on (CDLI)), framework (repository to make notifications about the project workflow), pyoracc (Python tools for working with Oracc) as well as mtaac_work and mtaac_gold_corpus software and the golden standard of MTAAC (Machine Translation and Automated Analysis of Cuneiform Languages). MTAAC is designed for the machine processing of a range of cuneiform languages that existed for quite a long time in Mesopotamia.

CDLI is written in seven languages (GitHub's inbuilt tool highlights Python, JavaScript, PHP and Java; it is also important to name Shell used for command line scripts).

The software is accessible through an open repository, however, it does not feature any thoroughly written readme files. A prospective user of CDLI will have to singlehandedly build each of the modules assisted only by his/her own coding skills in the seven programming languages mentioned earlier. However, it is possible to build this software on one's own.

Lingvodoc

The aim of Lingvodoc is to support the accumulation of data about different languages and dialects. It is a follow-up to the earlier project Dialeqt (Lingvodoc Repository). The project website (Lingvodoc) features a comprehensive database of research in languages and dialects of different nations living, mostly, in Eurasia. The biggest part of database sources is presented in the form of a tree that reflects the genetic kinship of languages. The search is based on a set of linguistic parameters. All the available sources, both corpora-based and lexicographical, are searchable. The project also features an interactive map and has a desktop application downloadable either through a conventional installer or a source code for the consecutive set-up.

Lingvodoc is open source and has its own GitHub repository (Lingvodoc Repository). Like CDLI, it is written in more than one programming language. Here, the major languages are JavaScript (the user database) and Python (processing within the database).

The software, as is said earlier, is available as a desktop application and a web version. The desktop application has a thorough manual on installation and updating as well as multiple functions allowing to create "the dictionaries of any structure", and "text corpora" (Lingvodoc). However, the programme has not been tested on all possible devices and has inevitable bugs fixing which will take a lot of effort of the developer and user alike. The web version, on the contrary, is a catalog, or an atlas of results that are already part of the database. It works correctly and is instrumental in the analysis of ancient languages. However, the web version does not feature an option of new corpora building.

Why not CDLI and/or Lingvodoc?

Both Lingvodoc and CDLI have some shortcomings that are crucial in the context of our research. They are not versatile and require special equipment which significantly limits their application. Besides, the sphere of CDLI application is narrow and CDLI itself is built from the source code.

The source code of the two projects might be used as a reference point in DIY software development initiatives, and/or as a source of the code fragments. To conclude, Lingvodoc and CDLI are quite effective in terms of enhancing a DIY product and contributing to the development of other projects.

Sketch Engine

The decision to use Sketch Engine (Sketch Engine) is one of the challenging choices for a researcher. The project itself has a great number of pros, however, it is pay walled. Sketch Engine has a thirty-day trial period. However, it is not enough to build a robust corpus. Pay wall is not a big issue; however, researchers may wish to opt for similar resources with limited functionality that are free. Such resources will be discussed in what follows.

The Sketch Engine website offers a subscription-based and a free trial access to the web application. The application allows to build a corpus in basically any language as long as the user can provide a link to the document in the network or a text file on their computer. Sketch Engine will automatically process all the received data. An automatic PoS tagging functionality is available for some languages. Unfortunately, the latter is not an option for either Ancient Greek or Old Church Slavonic. However, given the text files, it is possible to build corpora for both languages and obtain a large number of statistical data, for instance, bigram frequency score. Researchers have reported success stories of designing dictionaries based on the corpora built with Sketch Engine (Kilgarriff 2013).

Notably, Sketch Engine is not very good at solving very specific tasks, e. g., building a corpus for a specific language. Thus, any complex text preprocessing has to be made manually or with the help of other software. Preprocessing is necessary for the majority of languages — from Old Church Slavonic with dozens of texts to Arabic with millions of tokens (Alfaifi, Atwell 2013). In other words, Sketch Engine maybe a time and effort consuming solution for users with specific issues on their research agenda.

Access to the Sketch Engine code is blocked, it is a proprietary pay walled software. It can be neither transformed for one's needs, nor enhanced. The only way is to manually or automatically process data, and then upload the result directly into Sketch Engine.

Many researchers have already done this. A lot of well-known big corpora are made with the Sketch Engine Software (Sketch English). One of them is KSUCCA, a corpus of Classic Arabic mentioned above. There are studies that describe its development (Alrabiah, Al-Salman, Atwell 2013). The Sketch Engine website (KSUCCA) offers search options based on collocations, thesaurus, frequency list of words and n-grams, and isolated words. Researchers have developed a special tag set for Arabic to PoS tag the corpus. The tag set was included into the Sketch Engine standard library (MADA). Thus, KSUCCA is a corpus with multiple functionalities in the language that could be more or less effectively processed using Sketch Engine.

This case shows almost all the available functions of Sketch Engine in action: processing dozens of millions of tokens, convenient representation, practical outcome of the efforts of a research team who made the preprocessing for a specific language easier. However, when facing an unknown text format (CCMH), Sketch Engine, at best, clears this text from technical symbols and does not provide any interpretation. Such result is only satisfactory.

AntConc

AntConc is distributed under the free license that prohibits any sort of commercial use (Anthony 2019). The project has no open repositories. At the same time, the License Agreement strictly forbids any attempts to make a reverse engineering of the system. There are no ways to enhance the functionality of AntConc, be it the fixing of original shortcomings or an attempt to use specific language material. From this point of view, Sketch Engine and AntConc are identical. However, AntConc is distributed without additional fee for an extended version which may prompt a researcher to choose AntConct rather than Sketch Engine.

Now, we will only briefly focus on the AntConc project website: it hosts distributives to download the software as well as a list of related publications, the developer's CV and contacts.

AntConc is distributed as executable files (.exe for Windows, for instance) launched on a user computer. Installation is not required, and this makes it different from Lingvodoc and CDLI and similar to Sketch Engine. However, the fact that the user computer configuration can cause critical errors affecting the performance of the application makes AntConc a bigger challenge to use than Sketch Engine. Here, it is important to underscore relative versatility of the tool. We tested all the functionalities of AntConc on Windows 10 which the developer does not recommended to use and faced no error. Thus, individual computer set-up and parameters are not as crucial an issue as it may seem at the outset.

The issue of versatility vs. specialization is no less relevant for AntConc than it is for Sketch Engine. AntConc does not perform material preprocessing taking text files "as is". Additionally, it only processes .txt files and only locally. It does not upload and process XML files, JSON files, data tables, or files with separated values. It is a restriction made by the developer.

Unlike Sketch Engine, AntConc has a narrower range of functionalities. They include KWIC (Key Word In Context), graph plotting, file view, analysis of n-grams, collocations, list of words from the DIY corpus only, and list of words from the DIY corpus in comparison with the word list from a reference corpus. There is basically no PoS, syntactic or semantic tagging tools available "out of the box". Open access tools include frequency lists and lemma lists for a few high-resourced and well-studied European languages. This limits a possible application of AntConc and makes it seemingly ineffective when it comes to building tagged DIY corpora. The only alternative is to use AntConc as one of the tools in the researcher's toolbox, which may be highly undesirable.

Overall, it is difficult to evaluate the effectiveness of AntConc since it is difficult to find a corpus built in AntConc — a significant difference with Sketch Engine that boasts a comprehensive database. As a matter of experiment, we took a text from (CCMH) to build a small text corpus of Old Church Slavonic. The text chosen for the experiment was Codex Suprasliensis, the biggest text in the collection. According to AntConc, it has a few thousand tokens. The specific form of representation and lack of preprocessing returns the following result: "...jen& grEdy)i . (ada s&vE- 3037129 zana pokazati . a (adama (ot& (Oz& (otr@- 3037130 Siti . (i padenij..." (CCMH). That does not look like Old Church Slavonic. However, with the help of self-written software, partial preprocessing was made. To test the result, we used the Prague Fragments, a text in Czech Church Slavonic. This time the experiment resulted in the following: "...мъ сѧ емоу : ~ ~ СВѢТ- ИДЛЪНА : на розъсо іѣн(а) Ѣко ветъхы : і новы ходатаі : прѣдътеч- е !хвъ..." (TITUS). Here Old Church Slavonic is far more recognizable. Some words are yet hard to distinguish, but this is mostly due to the fragmentary nature of manuscripts, and some preprocessing flaws. In both cases AntConc was able to show lists of collocations, frequency lists of lexemes, and n-gram analysis. However, further text processing is again the responsibility of a researcher and the only possible solution is the development of additional software.

The experiment showed yet another drawback of AntConc. Unlike Sketch Engine that stores corpora on the project servers, storage and representation in AntConc is again the responsibility of a researcher who has to develop a special resource to store all the relevant data in the suitable format and view.

Despite the shortcomings mentioned above, AntConc is capable of performing the necessary minimal number of operations being, at the same time, free, accessible and easy-to-use which compensates for the lack of functionality. AntConc really comes in handy for simple tasks and as a blueprint. However, it may fail to be sufficient for certain applications.

#LancsBox v 5.0

#LancsBox (Brezina et al. 2020) was developed by a group of Lancaster University staff and students. It has lot in common with AntConc. In a way, it is an upgraded version of the latter. #LancsBox is a desktop application tested on Windows, MacOS, and the most popular Linux distributives. It is not available in open repositories and there is no indication of the programming language it uses.

#LancsBox stands out from its counterparts discussed earlier as it supports a bigger number of languages "out of the box", namely, English, Chinese and French by default, and approximately a dozen more accessible through downloading. This functionality is not provided automatically: to work with additional languages the user has to contact the developers. None of the extinct languages in the focus of archaeolinguistic studies was integrated in the system, which makes the product less promising for specific research tasks.

Nevertheless, #LancsBox is capable of conducting primitive lemmatization through fine-tuned tokenization. The test analysis of preprocessed Kiev Leaflets (TITUS) revealed 1,100 tokens attributed to 545 lemmata with the usage of pre-trained by developers lemmatizer for English. Russian lemmatizer discovered 0 lemmata, while picking "another language" prevented the programme from searching lemmata at all.

#LancsBox has a range of functions that are similar to those of AntConc.

KWIC proved to be quite effective even when working with texts in unknown languages that were preprocessed either manually or with additional software. The latter, however, makes the system a bit less attractive.

#LancsBox provides word frequency scoring, including more interpretable than raw frequency value of items per 10,000. The tokenization and token frequency scoring proved to perform effectively. Beside the frequency score, #LancsBox statistical functionality includes other metrics used to analyse different units such as a lemma, PoS or token.

Another #LancsBox vantage is an option to conduct basic search for collocates and analysis of n-gram occurrences in the text.

The difference between AntConc and #LancsBox is that #LancsBox visualizes collocations even in the corpora based on an unknown language. In addition, the user who uploads a text into the corpus can still see it. It might be an extra perk for some corpora, since the transfer of a text into the database may create some problems with the perception of its encoding by software.

#LancsBox, especially #LancsBox v. 5.0 in question, is a relatively new tool with only a few corpora built with its help. This makes it hard to find an example to illustrate its performance. To test the tool, we used #LancsBox to upload the preprocessed text of Kiev Leaflets (TITUS) mentioned above. The tool did not perform effectively enough to satisfy the needs of archaeolinguistics, however, it was able to perform basic analysis despite the lack of ancient languages in the database.

#LancsBox is one of the best solutions in its class. It provides wide possibilities for a big number of languages, and, with effective data preprocessing and pre-training of models, is capable of providing satisfactory results. Nonetheless, it does not allow to upgrade the existing functionality and does not provide an access to the code. Besides, it has a limited range of languages for "out of the box" analysis. All these factors prevent #LancsBox from being a number one choice among the corpora-building tools for archaeolinguistics.

Tsakorpus

When compared with the previous projects, Tsakorpus (Tsakorpus Repository) obviously presents a very special case. It is a completely open source software distributed by MIT License that puts basically no restrictions: a similar case was mentioned earlier (the open part of the CDLI code). This software is completely free, which is a big perk compared to the previously mentioned tools. Besides, the product itself might be enhanced, up to the point of being integrated into the user's own programmes. It is a significant step further in contrast with the previous platforms.

The first step to install Tsakorpus is to download its repository content, as is the case with some parts of CDLI. The application build is done through a terminal. The requirements include an installed Python interpreter and some very specific versions of Python libraries. It is recommended to run the build on the Apache server. All these requirements are neither difficult nor impossible to meet, however, the readme file lists possible hindrances. The project is a web app located on the user computer. This may lead to unforeseen mistakes connected with respective Python libraries. In this case, "out of the box" installation is impossible. Some OSs require to fix the source code. According to the developer of Tsakorpus, the version for Windows passed the test. Nonetheless, the latest Windows 10 builds witnessed significant difficulties in running the platform.

The highest degree of software versatility requires the most specific preparation for the specific corpus, specific research, and specific equipment. Tsakorpus is instrumental in building a corpus in basically any language, despite almost lacking "out of the box" solutions as compared, for example, with Sketch Engine. Tsakorpus is far closer to AntConc, but its modules allow adding files for different types of tagging, both morphological and syntactical. Tsakorpus and sufficient amount of time are what the research team needs to conduct a highly robust analysis of a small number of texts. Bigger projects may benefit from machine learning techniques, which is a viable option due to the programming language of the project.

Python has remarkable positive and negative sides relevant for the task in question. It is easy to learn, yet hard to master. It is slow, and if a software structure implies having a robust neural network within, it will take dozens of hours to execute. However, with most of the other programming languages, this time would be taken by writing one's own neural networks, or programs that utilize more basic machine learning techniques. Python offers numerous options for both these tasks and these options are much easier to find as compared with C++ or C#. Python can easily integrate codes in other languages so the researcher is free to work with the code as they want (excluding the necessary phase of finding integration libraries themselves).

Relative convenience and high versatility made Tsakorpus very popular as a corpora-building tool. The only tool that boasts a similar number of corpora built on its platform is Sketch Engine. The model of software distribution does not imply the existence of a single aggregator, yet the largest part of its corpora is available in (LCaS). The corpora feature many different languages, each possessing robust functionality.

An example we would like to discuss at this point is the Albanian corpus (Morozova, Rusakov, Arkhangelsky). The corpus is in two parts with modern language texts and old manuscripts, now presented as a single text. The corpus has both linguistic and metadata tagging. Metadata include the text name, the author (or a newspaper name), creation date and text genre/type. Linguistic information includes homophonic PoS tagging, glossing and translation of lemmas into English. Thus, it outstrips the examples provided earlier (excluding, maybe, only the most robust Sketch Engine ones). Besides, the corpus provides several search options based on grammatical categories, lemmatized forms, and even sentences. The search is as effective as in AntConc. This solution ideally fits the requirements of archaeolinguistic research.

At the moment, Tsakorpus seems to be the most effective tool. It is highly versatile, open source, and provides freedom of action for a linguist with sufficient technical expertise. However, some issues with installation may prevent some users from full-fledged integration of Tsakorpus into their product.

How exactly these projects might be of help?

Such products as Tsakorpus, #LancsBox, Sketch Engine, and AntConc give researchers certain freedom. They may be used as a prototype for a new system, or an important tool source. Sketch Engine, #LancsBox, and AntConc are mostly sources of inspiration, whereas Tsakorpus is basically a recipe for developing corpus-building software.

Conclusion

The article provided an overview of current state of research at the intersection of corpus linguistics and archaeolinguistics. The analysis revealed the lack of text corpora of ancient languages. Our aim was to consolidate information about the existing projects, including those that do not specialize in ancient languages, yet may find application in the development of respective tools.

First, we made an overview of existing ancient languages corpora. The analysis showed that tagged corpora of ancient languages are insufficient and the available corpora are not big enough. This necessitates the development of DIY corpora, including those based on the available ones.

This is where corpora building tools come in handy. We explored the projects that provide out-of-the-box, yet restricted, functionality, as well as those that allow a researcher to actively customize the solution uploaded into their system. Most corpora-building tools were found insufficient for the tasks facing archaeolinguistics. At the same time, some tools such as Tsakorpus and MTAAC may find effective application in the development and implementation of DIY software.

The promise for the follow-up study lies in the field of applied research, namely, in the development of our own Old Church Slavonic corpus and the versatile multifunctional software free from the shortcoming of earlier systems that had a negative impact on user experience.

Sources

Anthony, L. (2019) *AntConc (Version 3.5.8)*. [Computer Software]. Tokyo: Waseda University. Available at: https://www.laurenceanthony.net/software (accessed 22.03.2020). (In English)

Brezina, V., Weill-Tessier, P., McEnery, A. (2020) *#LancsBox v. 5.x*. [Computer Software]. Lancaster University. Available at: http://corpora.lancs.ac.uk/lancsbox (accessed 21.03.2021). (In English)

CCMH — *Corpus Cyrillo-Methodianum Helsingiense*. [Online]. Available at: https://korp.csc.fi/download/ccmh-src (accessed 24.02.2020). (In English)

CDLI — *The Cuneiform Digital Library Initiative*. [Online]. Available at: https://cdli.ucla.edu/ (accessed 08.03.2020). (In English)

CDLI Core Update — CDLI Core Update. *The Cuneiform Digital Library Initiative*. [Online]. Available at: https://cdli.ucla.edu/?q=news/cdli-core-update (accessed 08.03.2020). (In English)

CDLI Repository — CDLI. *GitHub Repository*. [Online]. Available at: https://github.com/cdli-gh (accessed 08.03.2020). (In English)

DDBDP — Duke Databank of Documentary Payri. *Papyri.info* [Online]. Available at: http://papyri.info/ddbdp (accessed 10.10.2020). (In English)

KSUCCA: King Saud University Corpus of Classical Arabic. *Sketch Engine*. [Online]. Available at: https://www.sketchengine.eu/corpus-of-classical-arabic-ksucca/#toggle-id-1 (accessed 21.03.2020). (In English)

LCaS — *Corpora and tools. Corpora of Russian Federation*. [Online]. Available at: http://web-corpora.net/?l=en (accessed 23.03.2020). (In English)

Lingvodoc 3.0. [Online]. Available at: http://lingvodoc.ispras.ru/ (accessed 10.03.2020). (In English)

Lingvodoc Repository — Lingvodoc repository on GitHub. *GitHub*. [Online]. Available at: https://github.com/ispras/lingvodoc (accessed 10.03.2020). (In English)

Manuscript. Slavyanskoe pis'mennoe nasledie [Manuscript. Slavonic written heritage]. [Online]. Available at: http://manuscripts.ru/ (accessed 24.02.2020). (In Russian)

Albanian National Corpus. (2016) [Online]. Available at: albanian.web-corpora.net (accessed 23.03.2020) (In English)

MTAAC — MTAAC Work Packages Repository. [Online]. Available at: https://github.com/cdli-gh/mtaac_work (accessed 08.03.2020). (In English)

Obshtezhitie — *The World Wide Web portal for the study of Cyrillic and Glagolitic manuscripts and early printed books*. (2020) [Online]. Available at: http://www.obshtezhitie.net/ (accessed 24.02.2020). (In English)

Perseus — PerseusDL/treebank_data. *GitHub*. [Online]. Available at: https://github.com/PerseusDL/treebank_data (accessed 24.02.2020). (In English)

PROIEL — Haug, D., Jøhndal, M. (2008) Creating a parallel treebank of the Old Indo-European Bible translations. In: C. Sporleder, K. Ribarov (eds.). *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Marrakech: European Language Resources Association Publ., pp. 27–34. (In English)

RRuDi — *A Russian Diachronic Online Corpus*. [Online]. Available at: https://www.slawistik.hu-berlin.de/de/member/meyerrol/subjekte/rrudi (accessed 24.02.2020). (In German)

SBLGNT — *SBL Greek New Testament*. [Online]. Available at: http://sblgnt.com/ (accessed 25.02.2020). (In English)

Sketch Engine — Text corpora in Sketch Engine. *Sketch Engine*. [Online]. Available at: https://www.sketchengine.eu/corpora-and-languages/corpus-list/ (accessed 21.03.2020). (In English)

Sketch English — Learn how language works. *Sketch English*. [Online]. Available at: https://www.sketchengine.eu/ (accessed 20.03.2020). (In English)

Tauber, J. K. (2017) *MorphGNT*: SBLGNT Edition. Version 6.12. [Online]. Available at: https://github.com/morphgnt/sblgnt (accessed 21.03.2021). (In English)

TITUS — *Thesaurus Indogermanischer Text- und Sprachmaterialien*. [Online]. Available at: http://titus.uni-frankfurt.de/indexe.htm (accessed 24.02.2020). (In English)

Tsakorpus Repository — Tsakorpus 2.0. *GitHub*. [Online]. Available at: https://github.com/timarkh/tsakorpus (accessed 23.03.2020). (In English)

USC OSC — *University of South California Old Slavic Corpus*. [Resource no longer accessible]. Available at: https://bcf.usc.edu/~pancheva/HistoricalSyntaxSouthSlavic.html#participants (accessed 24.02.2020). (In English)

Zeman, D., Nivre, J., Abrams, M. et al. (2020) *Universal Dependencies 2.6, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. [Online]. Available at: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3226 (accessed 21.03.2021). (In English)

References

Alfaifi, A., Atwell, E. (2013) Arabic Learner Corpus: Texts transcription and files format. In: *Proceedings of the International Conference on Corpus Linguistics (CORPORA-2013)*. Saint Petersburg: Saint Petersburg University Press, pp. 1–8. https://www.doi.org/10.13140/2.1.3468.8963 (In English)

Alrabiah, M., Al-Salman, A., Atwell, E. S. (2013) The design and construction of the 50 million words KSUCCA. In: *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*. Leeds: The University of Leeds Publ., pp. 5–8. (In English)

Alrabiah, M., Al-Salman, A., Atwell, E. S. et al. (2014) KSUCCA: A key to exploring Arabic historical linguistics. *International Journal of Computational Linguistics (IJCL)*, 5 (2): 27–36. (In English)

Arkhangelsky, T. A., Kisilier, M. L. (2018) Korpusa grecheskogo yazyka: dostizheniya, tseli i zadachi [Corpora of modern Greek: Achievements and goals]. In: N. N. Kazansky (ed.). *Indoevropejskoe yazykoznanie i klassicheskaya filologiya — XXII (chteniya pamyati I. M. Tronskogo). Materialy Mezhdunarodnoj konferentsii, prokhodivshej 18–20 iyunya 2018 g. [Indo-European linguistics and classical philology (Joseph M. Tronsky memorial Conference). Proceedings of the International Conference, St. Petersburg, 18–20 June, 2018]. Pt 1.* Saint Petersburg: Nauka Publ., pp. 50–59. https://www.doi.org/10.30842/ielcp230690152203 (In Russian)

Berkowitz, L., Johnson, W. H. (1990) *Thesaurus Linguae Graecae Canon of Greek authors and works*. 3rd ed. New York: Oxford University Press, 536 p. (In English)

Dandapat, S., Sarkar, S., Basu, A. (2004) A hybrid model for Part-of- Speech Tagging and its application to Bengali. In: *Proceedings of the International Conference on Computational Intelligence, ICCI 2004*. Istanbul: Esenyurt Univercity Publ., pp. 169–172. (In English)

Eckhoff, H. M. (2018) A corpus approach to the history of Russian po delimitatives. *Diachronica*, 35 (3): 338–366. https://doi.org/10.1075/dia.00006.eck (In English)

Hasan, F., UzZaman, N., Khan, M. (2007) Comparison of different POS tagging techniques (n-gram, HMM and Brill's tagger) for Bangla. In: K. Elleithy (eds.). *Advances and innovations in systems. Computing sciences and software engineering*. Dordrecht: Springer Publ., pp. 121–126. https://doi.org/10.1007/978-1-4020-6264-3_23 (In English)

Kilgarriff, A. (2013) Using corpora as data sources for dictionaries. In: H. Jackson (ed.). *The Bloomsbury companion to lexicography*. London: Bloomsbury Publ., pp. 77–96. https://www.doi.org/10.5040/9781472541871.ch-006 (In English)

Kopotev, M. (2014) *Vvedenie v korpusnuyu lingvistiku*. Prague: Animedia Company Publ., 185 p. (In Russian)

MADA — Habash, N., Rambow, O., Roth, R. (2010) *MADA+TOKAN Manual*. [Online]. Available at: http://www1.cs.columbia.edu/~rambow/software-downloads/CCLS-10-01.pdf (accessed 21.03.2020). (In English)

Mitrenina, O. (2014) The Corpora of Old and Middle Russian texts as an advanced tool for exploring an extinguished language. *Scribum*, 10 (1): 455–461. (In English)

Molina, M., Molin, A. (2016) In a lacuna: Building a Syntactically annotated corpus for a dead cuneiform language (on the basis of Hittite). In: *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2016". (Moscow, June 1–4, 2016)*. Moscow: Russian State University for the Humanities Publ. [Online]. Available at: http://www.dialog-21.ru/media/3476/molinammolina.pdf (accessed 21.03.2021). (In English)

Sokolov, E. G. (2019) The project of a deeply tagged parallel corpus of Middle Russian translations from Latin. *Journal of Applied Linguistics and Lexicography*, 1 (2): 337–364. https://www.doi.org/10.33910/2687-0215-2019-1-2-337-364 (In English)

Vendina, T. I. (2002) *Srednevekovyj chelovek v zerkale staroslavyanskogo yazyka***.** Moscow: Indrik Publ., 336 p. (In Russian)

Zakharov, V. P. (2015) Istoricheskie korpusa i korpusnye diakhronicheskie issledovaniya. In: *Pis'mennoe nasledie i informatsionnye tekhnologii "El'Manuscript-2015"*. Novosibirsk: State Public Scientific-Technological Library of the Siberian Branch of the RAS Publ., pp. 11–13. (In Russian)

Author

Ilia Afanasev, SPIN: 1382-6328, e-mail: szrnamerg@gmail.com

For citation: Afanasev, I. (2020) A corpus-based approach in archaeolinguistics. *Journal of Applied Linguistics and Lexicography*, 2 (2): 147–159. https://www.doi.org/10.33910/2687-0215-2020-2-2-147-159

Received 24 March 2021; reviewed 14 April 2021; accepted 14 April 2021.

Copyright: © The Author (2020). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.