УДК 81'33 doi: 10.22250/24107190_2023_9_4_35

Dmitry Yu. Gruzdev, Dmitry O. Kodzhebash
Military University, Russian Ministry of Defense
Moscow, Russian Federation
gru@inbox.ru

POS-powered queries for neat and lean concordances in ad-hoc corpora analysis

Abstract
The paper addresses the ways of boosting the query effectiveness in ad-hoc, a.k.a. DIY corpora. Based on the established trend of abandoning KWIC approaches in favor of pattern searches in the practice of professional translators, we hypothesize that advanced methods, including regular expressions and annotation, have a potential of accommodating them. The former has already proven an efficient text analysis tool for programmers while the latter takes credit for today's NLP technologies. It is the aspect that can be harnessed for increasing the efficiency of performing translation tasks. The obvious problem with ad-hoc corpora is that they rarely mature beyond the raw corpus status. However, tagging automation makes further improvements desirable. To grasp the potential of a tagged DIY corpus, we picked a random collection of news texts published in 2019–2020 and subjected it to automatic POS-tagging, acting as a computing lingua franca for more comprehensive and less constraining queries. The primary goal was to retrieve patterns based on the parts of speech of constituent words rather than anchor the search efforts to specific words. Taking advantage of existing tools, namely AntConc and TagAnt, we test the hypothesis. This confirmed the relevance of POS tagging of ad-hoc corpora for template extraction of linguistic data by translators. The resulting concordances, more concise and relevant, go a long way in keeping the analysis time lower, hence better overall efficiency.

Keywords: tags, POS, regular expressions, information retrieval, query, annotated corpus

For citation: Gruzdev, D. Yu., Kodzhebash, D. O. (2023). POS-powered queries for neat and lean concordances in ad-hoc corpora analysis. *Teoreticheskaya i prikladnaya lingvistika* [Theoretical and Applied Linguistics], 9 (4), 35–48. https://doi.org/10.22250/24107190_2023_9_4_35

Груздев Дмитрий Юрьевич, Коджебаш Дмитрий Олегович
Военный университет
г. Москва, Российская Федерация
gru@inbox.ru

Поиск лингвистической информации в аннотированном электронном корпусе текстов

Аннотация
В статье рассматриваются способы повышения эффективности запросов в специальных корпусах текстов. Исходя из сложившейся в практике профессиональных переводчиков тенденции отказа от поиска по ключевому слову в пользу шаблонных запросов, выдвигаются гипотеза, что регулярные выражения и аннотация, поддерживаемые современными корпус-менеджерами, имеют потенциал для более продуктивного извлечения лингвистической информации из корпуса текстов. Регулярные выражения уже полноценно используются программистами для анализа текста, в то время как аннотация легла в основу всех современных технологий обработки естественного языка. Это может быть использовано для повышения эффективности решения переводческих задач. Очевидной проблемой специальных корпусов является узкое окно релевантности, в результате чего скорость составления и подготовки ресурсов становятся главными критериями наряду с репрезентативностью. Это соответствует требованиям переводчиков,

постоянно сталкивающихся с потоком текстов из различных областей, но в то же время препятствует дальнейшему совершенствованию подготовленных корпусов текстов. Таким образом, жизненный цикл большинства специализированных корпусов завершается поиском и аккумуляцией текстов в соответствии со стоящей лингвистической задачей. Однако благодаря автоматизации аннотации дальнейшее совершенствование ресурсов становится перспективным. Чтобы оценить потенциал аннотированного специализированного корпуса, мы провели разметку корпуса новостных текстов за 2019–2020 гг. С помощью морфологической разметки был сгенерирован промежуточный язык между корпус-менеджером и корпусом текстов для реализации сбалансированных запросов для извлечения конструкций и фраз на основе частей речи без привязки к конкретным лексическим единицам. Для проверки гипотезы использовались корпус-менеджер AntConc и программа для морфологической разметки TagAnt. В результате подтвердился потенциал морфологической разметки по подготовке специальных корпусов для шаблонного извлечения лингвистических данных при поиске путей преодоления переводческих трудностей. Составленные конкордансы отличаются более ёмкими и релевантными совпадениями, что способствуют сокращению времени анализа.

Ключевые слова: теги, частеречная разметка, регулярные выражения, извлечение лингвистической информации, запрос, аннотированный корпус

Для цитирования: Gruzdev D. Yu., Kodzhebash D. O. POS-powered queries for neat and lean concordances in ad-hoc corpora analysis // Теоретическая и прикладная лингвистика. 2023. Вып. 9, № 4. С. 35–48. https://doi.org/10.22250/24107190_2023_9_4_35

1. Introduction

In the past five decades, corpora have become key to multiple NLP solutions. However, text selection, which was extremely time consuming when computer technologies were in their infancy, went as far as providing manual search and data retrieval. A true multiplier of corpus studies came in the form of annotation, tagging and parsing. Today's state-of-the-art language technologies – speech-to-text (STT), voice synthesis and recognition, machine translation, etc. – are all products of annotated corpus in one way or another [Hlaing et al., 2022]. For example, the naturalness of speech output in voice synthesis is as much dependent on phonemic representations as lexical stress markings, syllable boundaries or part-of-speech tagging [Lőrincz et al., 2021].

A significant portion of modern efforts in the development of e-tools for intercultural communications is marked by the departure from statistics in favor of semantics, which is covered by annotation as well. This is also relevant for application within a single language, for example, to extract a summary "based on a combination of semantics and statistics [Widyassari et al., 2022]." Other noteworthy example is the development of an adaptive information extraction system in biomedical domain based on high-quality semantically annotated corpora [Roberts et al., 2009].

Annotation has become an integral part of modern corpora used in advancing AI language solutions. The trend takes end users out of the loop, turning the technology into a black box. No matter the results, impressive by any standards, the lack of understanding takes the initiative of corpus application and adjustment from the hands of professionals.

Under the circumstanced any ad-hoc solution needed the involvement of a programmer. Not that there have been few in the past decade. Most were done in under resourced languages or to meet requirements of people occupied in fields other than linguistics. Again, non-linguists find annotated corpora useful for automatic information retrieval. For instance, some recent studies focused on making computers distinguish multitoken titles of entities or names of proteins in medical researches [Yamamoto et al., 2005].

Back to linguists, the feature of choice still remains the KWIC (key word in context) search option. This is accounted for by the fact that people in this trade, translators in particular, tend to use words or their combination to overcome difficulties and find a solution to a problem. However, as their language proficiency grows, translators gain everything there is to engage corpora in checking grammar structures [Zanettin, 2013]. However, the use of specific words will prove to be counterproductive due to the lack of fuzzy search algorithm in corpus-managers.

Template search in a corpus of texts allows to make up for the lack of the algorithm. In this respect regular expressions have become a major tool providing a flexible context of the checked lexical unit, thus increasing the efficiency of the linguistic information retrieval [Gruzdev et al., 2022]. The main drawback of the corpus is that the computer perceives a word as a chain of characters and the text as a sequence of character groups [Albukhitan et al., 2020]. Therefore, when writing a query pattern, it is essential to integrate wild cards, for example, with formal attributes of parts of speech (POS) to clarify the role of the searched unit context. This way the program is instructed as to the functions of character groups in a sentence. Annotation on the other hand can specify additional features for each word in advance, thereby preparing a large array of texts for pattern analysis. Due to the labor-intensive processing, i.e. tagging and annotation, in corpus linguistics, the approach has not been used by translators until recently. The emergence of automatic corpus markup software calls for evaluation of the effectiveness of regular expression (RegEx)-powered template search in the annotated corpus.

2. Tools

To achieve the goal set in the paper, the research will need the following tools and instruments: 1) corpus manager supporting regular expressions, 2) automatic annotation tools, 3) additional processing tools for optimizing text parameters in the corpus, 4) news corpus of texts in .txt format.

When selecting resources, specific attention was paid to the functionality of the software, which should include a variety of search settings, support regular expressions, and annotated corpus queries. In previous studies, the preference was given to the AntConc corpus manager packing all the necessary functionality. Due to the generated experience with the software, its accessibility and user-friendly interface, the choice was made in favor of AntConc 3.5.9 [AntConc, 2020]. In addition to the corpus manager, its designer Anthony Laurence developed a number of corpus research software, including automatic annotation software TagAnt and file preparation software AntFileSplitter, which meet the requirements of this study as well [TagAnt, 2022; AntFileSplitter, 2019].

For a raw corpus, it was decided to adopt 2019 and 2020 news corpora available online [Goldhahn et al., 2012]. Note that the qualitative aspect of the corpus does not constitute a factor in the attainment of the goal, thus the fact of having a ready-made tool was decisive. The sample of news texts published within a certain period meets the basic requirements for this class of electronic linguistic resources, namely – consistency in time and topics, and representativeness within the framework of the study [Gruzdev, Gruzdeva & Makarenko, 2019].

In accordance with the goal of the research, additional processing of the corpus was carried out with TagAnt followed by the breaking of the bulk files of the original corpus to the optimal size by AntFileSplitter to increase the AntConc processing speed.

The resulting tool will be used to conduct a series of experiments to evaluate the effectiveness of RegEx-powered patterned information retrieval strategies in an annotated corpus.

3. Materials and methods

With the development of electronic linguistic corpora in the second half of the 20th century it became possible to organize the storage and processing of large arrays of texts.

However, to automate the analysis and extraction of information, it was necessary to train computers to understand natural language. As a result, efforts were mounted to standardize language interpretation. By the end of the 1990s, a terminological basis had taken shape: a distinction was made between markup and annotation. The former provides extralinguistic characteristics of the language, for example, the genre of the text, situation, gender, age of the participants in the dialogue, etc. As a rule, these do not lead to ambiguity. The opposite happens in annotation whose purpose is to interpret the language. In this case, the potential for ambiguous results is much greater [Gruzdev, Gruzdeva & Makarenko, 2019]. It is this area of corpus processing that paved the way to breakthrough technologies in language application in the early 2000s and not so much time later proved to be a stumbling block.

The new millennium became a watershed for the corpus linguistics. During its conception and development, most of the projects were manual, garnering the interest of specialists. Everyone could try to adapt the resource to their own needs. Thus, translators saw enormous opportunities in it to search for linguistic information, to test their solutions in terms of quality and efficiency, and even relevance in the target language – something that had been achievable in a conversation with a native speaker only [Gruzdev et al., 2019].

However, advances in the first five years of the 2000s led to the automation of the translation process and the design of sophisticated products to meet the needs of most potential users. There was a leap from the study of the annotation problem to the successful application of the results in automated systems. People owe this to the emergence of voice recognition and speech transcription software, machine translation, and search engines with fuzzy search capabilities [Gruzdev et al., 2019]. At the same time, the interests of narrow specialists, including translators, were left unabated. These tools no longer seemed like raw resources. Since 2006, certain open-ended projects in the field of text processing automation, among them the AOT initiative by the Russian State University for the Humanities, have fallen into oblivion [Zanettin, 2013].

The annotated corpus is still used in individual language studies, for example, when compiling dictionaries and conducting special projects [Zanettin, 2013]. With the integration of an interlingua in the form of annotation, it is possible to highlight to the computer any features of the texts being studied in the corpus. Thus, completely exotic projects have emerged, one of them being S. Granger's educational annotated corpus compiled in 1998 to include papers of English language learners. S. Granger deliberately marked up the errors [Laviosa, 2004].

Translators are more likely to use specialized corpora without additional processing in the form of markup and annotation. The reason for this is the labor-intensive nature of the additional preparation of a corpus, which is unacceptable given the permanent backlog and ever-shifting subject matters of texts coming for translation [Soumia et al., 2017]. Furthermore, as a result of the logical decline in the need for the tool as more experience is gained in a certain area makes the prospect of additional hours or days spent improving the resource less appealing.

The solution to the problem lies in automation of the annotation stage. Since the purity, relevance and representation of the ad-hoc or DIY (do-it-yourself) corpus determines the quality of the resource, the translator's participation in the selection of material is essential [Gruzdev, 2011]. As experience is gained, the time to compile a raw corpus can be reduced to tens of minutes. However, further processing of the prepared instrument requires considerable time and the old man-in-the-loop approach seems as essential as ever before. What has not been said in the paper so far is that the breakthrough in the early 2000s resulted precisely from the automation of standard types of annotation – albeit with a slight reduction in quality from 98% to 95–96% – which can accelerate the process of extracting linguistic information from electronic corpora of texts [Névéol et al., 2010].

The standard types of annotation correlate with the main levels of language. A distinction is made between POS (part-of-speech), syntactic, prosodic, phonetic, and

semantic annotation. The type, the depth and quality of processing determine the applied value of the instrument. It is essential for a translator to find information quickly, so query and search will be decisive in selecting the type of annotation used.

The main challenge is to standardize natural language for the computer. The logic is simple: the computer perceives text as a continuous series of groups of characters that in natural language humans perceive as words. Similar to mathematical examples, variables must be introduced to group sentences into patterns and provide pattern information retrieval. It should be noted that programmers have achieved standardization of language by using regular expressions, which have also been successfully used to extract linguistic information from a corpus [Gruzdev et al., 2019]. However, programmers deal with ordered syntax as opposed to natural language with all the variety and unpredictability that the translator encounters in practice.

For modeling pattern-based queries, it is essential to establish recurring relationships in a language. Naturally, these are observed at the sentence level. An optimal annotation would be a mask breaking the text down into sentence members, which is difficult to perform automatically without further disambiguation due to the lack of persistent features in the groups of characters pertinent to the sentence parts [Albukhitan et al., 2020]. In addition, the same lexeme can perform the functions of different sentence members, adding to the ambiguity, thus further complicating querying. In view of all the limitations, it is advisable to consider the text for regular and recurring patterns, e.g., at the level of parts of speech.

At the POS level, annotation can be performed by automatic means with minimal errors. Moreover, this type of corpora processing is the most common and elaborated essentially laying the groundwork for other standard annotations [Alhasan & Al-Taani, 2018]. Therefore, it will only be logical to study POS annotation more thoroughly for mating it with the RegEx function of corpus managers to boost the efficiency of search and retrieval of linguistic information.

4. Results

In any annotation a set of tags is of primary importance. It is this aspect that shapes the intermediate language that ensures the automation of the corpus analysis and querying [Alkhatib et al, 2021; Shrestha & Dhakal, 2021]. Most standard annotations are individualized, which does not provide continuity in researches and complicates future refinements and improvements. From this point of view, POS annotation is the most standardized in terms of the tag set [Gruzdev et al., 2022].

Indeed, there is little variability, but the basis is always one of the CLAWS (Constituent Likelihood Automatic Word-tagging System) sets. Needless to say that even this successful model has not overcome the plague of all annotation and markup systems, i.e. ambiguity which remains at 3%. The project developed eight sets of C1-C8 tags varying in depth and thoroughness. It was assumed that C8 with the most elaborate set would be able to distinguish particular POS instances. However, this also failed to eradicate the problem. As it turned out, more tags lead to a higher degree of ambiguity in the annotation process but allow a more precise querying in processed corpora. Thus, the most common C5 (60 tags) and C7 (152 tags) were used in the British National Corpus (see Example 1, CLAWS). The smaller set of tags was used to process the entire corpus while the more elaborate one was applied to a small sample of two million tokens followed by manual verification.

Given the challenges and needs of a translator to compile an annotated corpus in the shortest possible time without having to disambiguate the results, it makes sense to limit yourself to a small set of tags. This is exactly what the developer of TagAnt, Anthony Lawrence, has done [TagAnt, 2022]. The software designer himself recommends the TreeTagger set of 58 tags (see Table 1, Treetagger Tag Set). In quantitative and qualitative

composition, the set is very close to C5. The success of this set in a large BNC annotation project proves its feasibility beyond any doubt.

Example 1. Excerpt from the British National Corpus, annotated with a set of C5 tags (from Captain Pugwash and the Huge Reward, British National Corpus)

<s c="0000002 002" n=00001>
When&AVQ-CJS; Captain&NP0; Pugwash&NP0; retires&VVZ; from&PRP;
active&AJ0; piracy&NN1; he&PNP; is&VBZ; amazed&AJ0-VVN; and&CJC;
delighted&AJ0-VVN; to&TO0; be&VBI; offered&VVN; a&AT0; Huge&AJ0;
Reward&NN1; for&PRP; what&DTQ; seems&VVZ; to&TO0; be&VBI; a&AT0;
simple&AJ0; task&NN1;.&PUN;
<s c="0000005 022" n=00002>
Little&DT0; does&VDZ; he&PNP; realise&VVI; what&DTQ; villainy&NN1;
and&CJC; treachery&NN1; lurk&NN1-VVB; in&PRP; the&AT0; little&AJ0;
town&NN1; of&PRF; Sinkport&NN1-NP0;,&PUN; or&CJC; what&DTQ; a&AT0;
hideous&AJ0; fate&NN1; may&VM0; await&VVI; him&PNP; there&AV0;.&PUN;

Table 1. **TreeTagger tag set recommended by TagAnt developer Anthony Lawrence (58 tags)**

POS Tag	Description	Example	POS Tag	Description	Example
CC	coordinating conjunction	*and, but, or, &*	**VB**	verb *be*, base form	*be*
CD	cardinal number	*1, three*	**VBD**	verb *be*, past	*was|were*
DT	determiner	*the*	**VBG**	verb *be*, gerund/participle	*being*
EX	existential there	*there is*	**VBN**	verb *be*, past participle	*been*
FW	foreign word	*d'œuvre*	**VBZ**	verb *be*, pres, 3rd p. sing	*is*
IN	preposition/subord. conj.	*in,of,like,after, whether*	**VBP**	verb *be*, pres non-3rd p.	*am|are*
IN/that	complementizer	*that*	**VD**	verb *do*, base form	*do*
JJ	adjective	*green*	**VDD**	verb *do*, past	*did*
JJR	adjective, comparative	*greener*	**VDG**	verb *do* gerund/participle	*doing*
JJS	adjective, superlative	*greenest*	**VDN**	verb *do*, past participle	*done*
LS	list marker	*(1),*	**VDZ**	verb *do*, pres, 3rd per.sing	*does*
MD	modal	*could, will*	**VDP**	verb *do*, pres, non-3rd per.	*do*
NN	noun, singular or mass	*table*	**VH**	verb *have*, base form	*have*
NNS	noun plural	*tables*	**VHD**	verb *have*, past	*had*
NP	proper noun, singular	*John*	**VHG**	verb *have*, gerund/participle	*having*
NPS	proper noun, plural	*Vikings*	**VHN**	verb *have*, past participle	*had*
PDT	predeterminer	*both the boys*	**VHZ**	verb *have*, pres 3rd per.sing	*has*
POS	possessive ending	*friend's*	**VHP**	verb *have*, pres non-3rd per.	*have*
PP	personal pronoun	*I, he, it*	**VV**	verb, base form	*take*

Continuation of Table 1

POS Tag	Description	Example	POS Tag	Description	Example
PP$	possessive pronoun	*my, his*	**VVD**	verb, past tense	*took*
RB	adverb	*however, usually, here, not*	**VVG**	verb, gerund/participle	*taking*
RBR	adverb, comparative	*better*	**VVN**	verb, past participle	*taken*
RBS	adverb, superlative	*best*	**VVP**	verb, present, non-3rd p.	*take*
RP	particle	*give up*	**VVZ**	verb, present 3d p. sing.	*takes*
SENT	end punctuation	*?, !, .*	**WDT**	wh-determiner	*which*
SYM	symbol	*@, +, *, ^, |, =*	**WP**	wh-pronoun	*who, what*
TO	to	*to go, to him*	**WP$**	possessive wh-pronoun	*whose*
UH	interjection	*uhhuhhuhh*	**WRB**	wh-abverb	*where, when*
			:	general joiner	*;, -, --*
			$	currency symbol	*$, £*

Having checked the TreeTagger tag set against a sample of text annotated with TagAnt, it was established that several tags have been modified or are not used at all (see Table 2). For instance, Example 2 shows that proper nouns are labeled NNP rather than NP. Also, punctuation marks are not combined into a separate category but rather have a tag of each own.

Example 2. Excerpt from the 2019 and 2020 news corpus [Goldhahn, Eckart & Quasthoff, 2012], annotated with TagAnt [TagAnt, 2022]

Acclaimed_JJ editor_NN and_CC Oscar_NNP winning_VBG sound_NN designer_NN Walter_NNP Murch_NNP who_WP had_VBD worked_VBN on_IN everything_NN from_IN to_IN would_MD make_VB the_DT big_JJ step_NN up_RP to_IN the_DT director_NN s_POS chair_NN

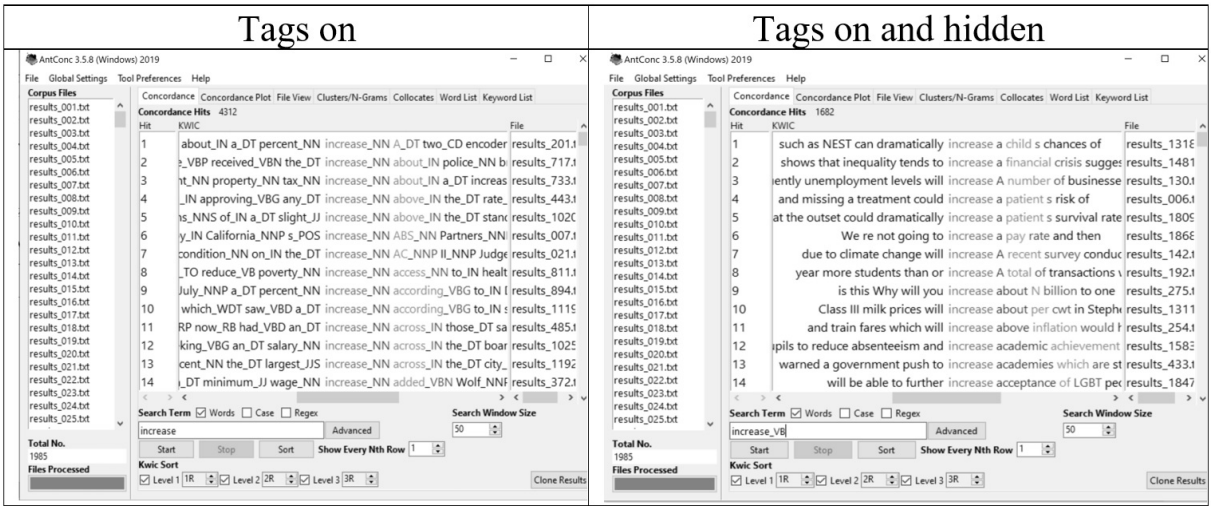
T a b l e 2. **Verified tag set (44 tags)**

POS Tag	Description	Example	POS Tag	Description	Example
CC	coordinating conjunction	*and, but, or, &*	**RBS**	adverb, superlative	*best*
CD	cardinal number	*1, three*	**RP**	particle	*give up*
DT	determiner	*the*	**SYM**	symbol	*@, +, *,л, 1, =*
EX	existential there	*there is*	**TO**	to	*to go, to him*
FW	foreign word	*d'xuvre*	**UH**	interjection	*uhhuhhuhh*
IN	preposition/subord. conj.	*In, of, like, after, whether*	**VB**	verb *be*, base form	*be*
IN/ that	complementizer	*that*	**VBD**	verb *be*, past	*was Iwere*
JJ	adjective	*green*	**VBG**	verb *be*, gerund/participle	*being*
JJR	adjective, comparative	*greener*	**VBN**	verb *be*, past participle	*been*

Continuation of Table 2

POS Tag	Description	Example	POS Tag	Description	Example
JJS	adjective, superlative	*greenest*	**VBZ**	verb *be*, pres, 3rd p. sing	*is*
LS	list marker	*(1),*	**VBP**	verb *be*, pres non-3rd p.	*am|are*
MD	modal	*could, will*	**WDT**	wh-determiner	*which*
NN	noun, singular or mass	*table*	**WP**	wh-pronoun	*who, what*
NNS	noun plural	*tables*	**WP$**	possessive wh-pronoun	*whose*
NNP	proper noun, singular	*John*	**WRB**	wh-abverb	*where, when*
NNPS	proper noun, plural	*Vikings*	**:**	general joiner	.. -. —
PDT	predeterminer	*both the boys*	**$**	currency symbol	*$, £*
POS	possessive ending	*friend's*	**.**		.
PRP	personal pronoun	*I, he, it*	**,**		,
PRP$	possessive pronoun	*my, his*	**“”**		“ ”
RB	adverb	*however, usually, here, not*	?		*?*
RBR	adverb, comparative	*better*	**HYPH**	-	rear-mounted

Now that TagAnt performed POS annotation of a corpus of texts automatically, the subsequent task is to test the effectiveness of the prepared tool. Extraction of linguistic information for overcoming translation difficulties was carried out in AntConc also developed by Anthony Lawrence [AntConc, 2019]. For the convenience of retrieving linguistic data from annotated corpora, the program has a function of hiding tags while retaining the ability to use them in queries. As a result, the concordance is not riddled by extra characters, violating the integrity of the text. In Figure 1 on the right, the tags are hidden, although the VB tag was used in the query to search for a verb rather than a noun.



F i g u r e 1. **Concordance retrieved from the annotated corpus with tags enabled and hidden**

5. Discussion

In its class, the raw corpus has the undeniable advantage of almost effortless preparation. However, in the absence of a fuzzy search algorithm, the user has to start searching for linguistic information with way too general, simple single-word queries. The approach calls for further refined queries based on primary long concordances with a large number of irrelevant matches. This may cycle on unloading a backlog of concordances on to the translator. The task gets more difficult when it comes to checking grammatical structures. In conditions of unknown lexical units, it is reasonable to develop the query based on marker words. The common nature of some lexical units leads to extensive concordances, thus increasing the time to find ways to overcome translation difficulties. For example, in the inverted structures in Table 3, frequent vocabulary makes up a significant part resulting in a dramatic drop in the relevance of the concordances. In some cases, it is not even possible to pick a marker. In the English structure So+ adjective + to be (so ...), the constant is a common adverb with a graphic spelling that coincides with the conjunction and the introductory word.

Table 3. **Inverted structures in English**

No. seq.	Word/Combination	No. seq.	Word/Combination	No. seq.	Word/Combination
1.	barely... when	10.	nor/neither	19.	only by
2.	hardly (ever)... when	11.	not (even) once	20.	only in this way
3.	in no way	12.	not only... (but) also	21.	only then
4.	in/under no circumstances	13.	not only do/will/can... as well	22.	rarely
5.	little	14.	not since	23.	scarcely (ever)... when
6.	little... know/realize	15.	until / not till	24.	seldom
7.	never	16.	nowhere	25.	So …
8.	never before	17.	on no account		
9.	no sooner... than	18.	on no occasion		

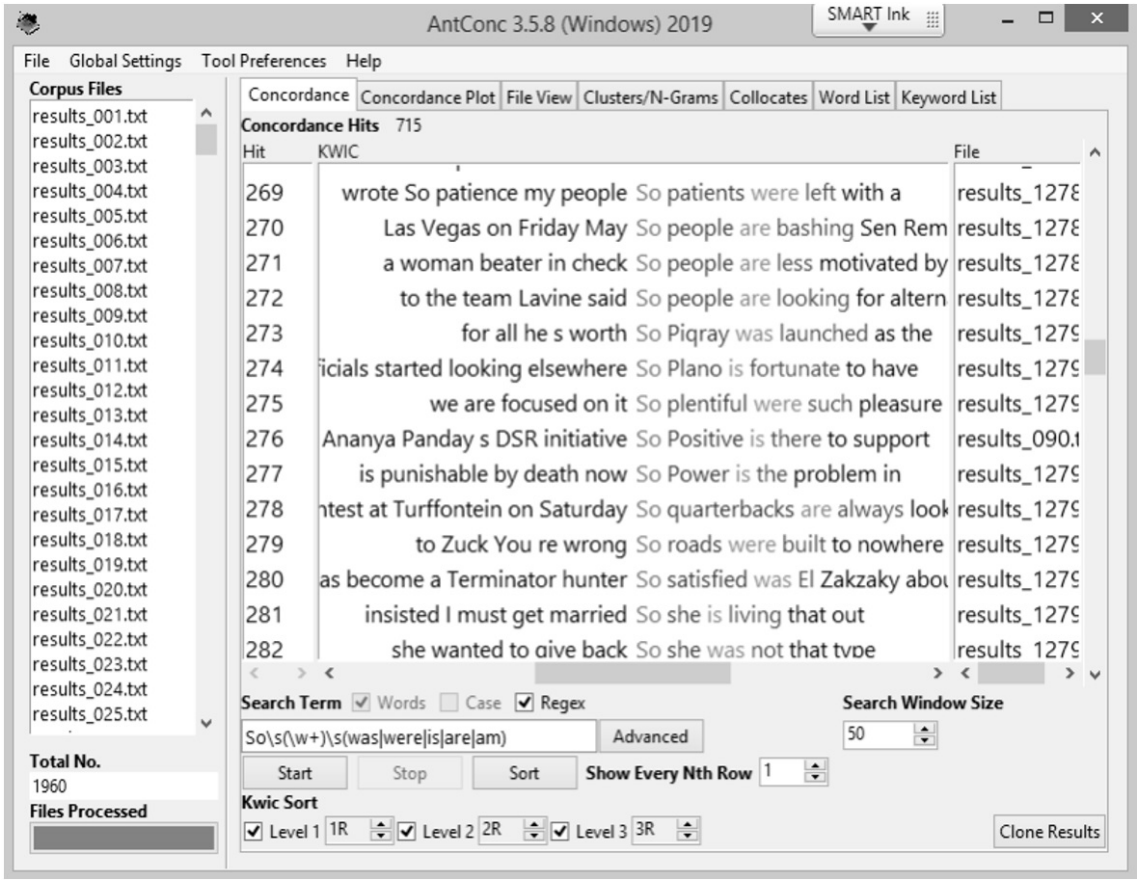
Checking of a sequence of two two-character sequences So in the corpus results in a concordance of 37,798 matches. Even a superficial analysis of such a volume would require at least 10 minutes with no guarantee of finding relevant hits.

One solution is to use the “regular expressions” function. Given the benefit of the tool to point to unknown words and set multiple choice for individual words in one query, one can refine the query without linking it to those lexical units whose probability of occurrence in the corpus in a particular sequence cannot be forecast in advance [Gruzdev et al., 2022].

Table 4. **Code components for pattern search of the inverse structure so+adjective+verb to be in a corpus of texts**

No.seq.	Task	Code
1.	Random adjective after *so*	(\w+)
2.	Probable form of the verb *to be*	(was|were|is|are|am)
3.	Space	\s

The query So\s(\w+)\s(was|were|is|are|am) resulted in 800 hits (see Table 4). After applying ABC sorting, it took below a minute to locate a relevant example (see Figure 2).



F i g u r e 2. **Results of the So\s(\w+)\s(was|were|is|are|am) query based on regular expressions**

It should be noted that even with a significant reduction of the concordance by a factor of 40 it was not possible to completely get rid of irrelevant hits: only 11 hits out of 800 examples were examples of the needed grammar pattern.

For the sake of comparison, it was decided to check the sequences So+ spontaneous adjective in the corpus (see Examples 3 and 4).

In Example 4, the case-sensitive feature was engaged to artificially narrow the search to cases at the beginning of a sentence. In the KWIC search free of additional functions, 139 matches were retrieved, none of which met the criteria of the searched phrase.

Given the low productivity, the verification of grammar structures in the corpus in the KWIC mode is inferior to the RegEx-powered method of linguistic information retrieval. However, the ratio of relevant queries to the total volume of the concordances provides no solid grounds to claim effectiveness of the second approach, either. Among the obvious problems one should note the lack of possibility to specify morphological features of the adjective. For example, if a verb was used in the gerund form, a solution would be to refine the query by appending the operator (\w+) with -ing. In the above example, the query had to indicate the presence of an indefinite word after the adverb So. Similarly, when it is necessary to indicate a number, regular expressions provide a combination (1–9) to denote an arbitrary series of digits [Gruzdev, Kodzhebash & Makarenko, 2022]. However, it is possible for numbers to be spelled out in the text, which this pattern will not recognize and will miss.

Given the rule of writing numbers up to 10 at the beginning of a sentence, the situation is not uncommon in English [Scribendi].

Example 3. A concordance extracted with AntConc from a news corpus of 1,000,000 words (query – so loud, 9 matches)

Hit No.	Concordance	
1.	the guy not to laugh	so loud I realised that it
2.	when the music isn t	so loud ok cutie Can t
3.	anything I feel like screaming	so loud right now it s
4.	the birds weren t quite	so loud Sarah Hammond who has
5.	Furious So why are they	so loud So why are we
6.	test hop and it was	so loud that All tankers need
7.	approximately three hours and was	so loud the walls were shaking
8.	black incomes So it is	so loud to play in So
9.	I ve laughed my heart	so loud Yes I do support

Example 4. A concordance extracted with AntConc from a news corpus of 1,000,000 words (query – so good with the case of the first letter of the adverb so, 8 matches)

Hit No.	Concordance	
1.	Goodell expected crap got it	So good for him for dropping
2.	t matter if you die	So good I thought for one
3.	him for dropping his gloves	So good in fact that it
4.	it motivated me to continue	So good it doesn t matter
5.	second about eating the floss	So good luck to them So
6.	So good luck to them	So good press is now more
7.	the start of the tournament	So good to be back in
8.	see Apollo back on screen	So good to see the children

These limitations are eliminated in the annotated corpus. Instead of the regular expression (1–9), the word symbol and the cardinal number tag (w+)_CD can be used to retrieve all references to quantities in the form of numbers and words from the corpus. Table 5 demonstrates the query for the template search of the inverted structure So+ adjective + to be (so ...) in the annotated corpus. In the query, regular expressions are specified by POS tags from Table 2. It extracted a concordance of 46 hits, 15 of which represent the grammatical phenomenon in question (see Example 5).

This concordance demonstrates that the program is case-sensitive in the RegEx mode. This fact should be considered as another aspect of computer's perception of a text. In the given example, the capital letter S in the query has bounded the search only to the beginning of sentences, which corresponds to the conditions of using the inverted structure So+ adjective + to be (so ...). Punctuation marks have a potential to offer another way to indicate where the

word or phrase in question shall occur in a sentence. However, the developers excluded them from the corpus employed for the purposes of this paper.

Table 5. **Code components for pattern search of the inverse construction so+adjective+verb to be in an annotated corpus of texts (request - So_RB\s(\w+)_JJ\s(\w+)_VB.\s)**

No.seq.	Task	Code
1.	Adverb *so*	So_RB
2.	Random adjective after *so*	(\w+)_JJ
3.	Random verb after the adjective	(\w+)_VB.
4.	Space	\s

Example 5. Matches from a concordance compiled with AntConc and a news corpus of 1,000,000 words (query - So_RB\s(\w+)_JJ\s(\w+)_VB. \s, 46 matches)

No.seq.	Nit No.	Concordance	
1.	2	arily struggling with that zipper	So bad has it been that
2.	3	of security among his people	So bad has the situation become
3.	5	in winning only one game	So catastrophic is the situation to
4.	6	or available as a Webinar	So few are surprised that as
5.	7	him into a better quarterback	So harsh was the sand on
6.	8	the community said Mulberry	So important was pietas to the
7.	37	to keep the YMCA open	So persistent was this narrative that
8.	38	wall but not a woman	So popular are the banned carrier
9.	39	history of the United States	So potent is the bigotry of
10.	40	Magneto to get her released	So profitable is the forex parallel
11.	41	carry us through to Jan	So serious is the situation that
12.	42	Matt Rudi Danny Bekah	So severe was this that six
13.	43	So training is very essential	So transformative was Gorbachev’s revolution
14.	44	that caused heavy damage	So vast is the wealth generated
15.	45	would you go anywhere else	So wild is on brand for

Another important observation made during the experiment is that tags also function according to the general rules of text analysis in a corpus. The program treats them as a sequence of characters, thus the RegEx wild cards are relevant in the POS tags as well. Most of the them are a sequence of 2–3 characters, and within the same category the first two characters are mostly the same. For example, all verbs and nouns begin with VB and NN respectively (see Table 2). In the query demonstrated in Table 5, in order to extend the concordance scope as far as verbs are concerned, the third character has been replaced by ".", which corresponds to null or any character in the corpus RegEx function [Gruzdev et al., 2022].

6. Conclusions

The analysis of strategies for retrieving linguistic information in ad-hoc corpora resulted in a number of significant conclusions. First, due to the emergence of freeware tools for corpus annotation, translators should consider upgrading raw DIY corpora to increase the efficiency of querying. Second, it is advisable to use the POS annotation which is the most elaborate and standardized to date. It will prove to be an asset in clarifying the query up to the functions and role of optional lexical units in a phrase or grammar structure without specifying their spelling. Thirdly, the combined use of the annotated corpus and regular expressions leads to a cumulative effect. As a result, the search time is reduced by a factor of ten. Fourth, when checking grammar structures in a corpus, it is counterproductive to stick to lexical units, therefore it has to be replaced by optional combination of regular expressions denoting a string of characters separated by spaces on either end, i.e. word, and POS tags specifying its role. Fifth, when building a RegEx query in an annotated corpus, one should not neglect secondary features in order to bind the structure or phrase in question to a specific place in the sentence. For example, a dot or a capital letter will narrow down the program's field of search to the beginning of sentences. Sixth, before using the annotated corpus, it is essential to clarify the set of tags and the extent of their implementation in texts. The exclusion of linguistic phenomena during the compilation of the corpus leads to a lack of grounds for the application of the appropriate tags. Seventh, although the TagAnt software used in the experiment has been endowed by an optimal set of features, it only supports seven languages. Thus, the next step in this field needs to focus on locating alternative tools to expand the language base. Finally, some types of annotations left out of the scope of this paper are also of interest from the point of view of stepping up the query efficiency. The study of semantic and syntactic annotations could be a logical follow on to this research.

References

Albukhitan, S., Alnazer, A., & Helmy T. (2020). Framework of Semantic Annotation of Arabic Document using Deep Learning. *Procedia Computer Science*, 170, 989–994. https://doi.org/10.1016/j.procs.2020.03.096

Alhasan, A., & Al-Taani, A. (2018). POS Tagging for Arabic Text Using Bee Colony Algorithm. *Procedia Computer Science*, 142, 158–165. https://doi.org/10.1016/j.procs.2018.10.471

Alshammari, N. and Alanazi, S. (2020). The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22. https://doi.org/10.1016/j.eij.2020.10.004

Anthony, L. (2022). *TagAnt* (2.0.5) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2020). *AntConc* (3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2019). *AntFileSplitter* (1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (May, 2023). *Treetagger Tag Set.* https://laurenceanthony.net/software/tagant/resources/treetagger_tagset.pdf

Garside, R. (May, 2023). *British National Corpus*. Department of Computing, University of Lancaster, UK. http://www.natcorp.ox.ac.uk/docs/garside_allc.html

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. *Proc. of the 8th International Language Resources and Evaluation* (LREC'12).

Gruzdev, D. (2011). Corpora as an interpreter's tool. *MSU Herald, Series 22, Translation Theory*, 2, 23–35. (In Russ.).

Gruzdev, D., Gruzdeva, L., & Makarenko, A. (2019). 'Regular expressions' as a way of dealing with translation difficulties. *MSU Herald, Series 22, Translation Theory,* 4, Retrieved October 20, 2022.

(In Russ.). https://cyberleninka.ru/article/n/preodolenie-perevod-cheskih-trudnostey-s-pomoschyu-regulyarnyh-vyrazheniy

Gruzdev, D. Kodzhebash, D., & Makarenko, A. (2022). Corpus query strategies. *Military Philological Journal*, 3, 77–91. (In Russ.).

Hlaing, Z., Kyaw, T., Supnithi, T., & Netisopakul, P. (2022). Improving neural machine translation with POS-tag features for low-resource language pairs. *Heliyon,* 8. https://doi.org/10.1016/j.heliyon.2022.e10375

Kotiurova, I. A. & Shchegoleva, L. V. (2022). Analysis of incorrect POS-tagging in student texts with linguistic errors in German. *Research Result. Theoretical and Applied Linguistics*, 8 (3), 87–99. (In Russ.). https://doi.org/10.18413/2313-8912-2022-8-3-0-6

Laviosa, S. (2004). Corpus-based Translation Studies: Where does it come from? Where is it going? *Language Matters*, 10, 29–31.

Lőrincz, B., Stan, A., & Giurgiu, M. (2021). An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis. *Procedia Computer Science*, 192, 756–765. https://doi.org/10.1016/j.procs.2021.08.078

Losey-León, María-Araceli. (2014). Corpus design and compilation process for the preparation of a bilingual glossary (English-Spanish) in the logistics and maritime transport field: LogisTRANS. *Procedia – Social and Behavioral Sciences*, 173. https://doi.org/10.1016/j.sbspro.2015.02.068

Nerabie, A. M., Alkhatib, M., Mathew, S., El Barachi, M., & Oroumchian, F. (2021). The Impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach. *Procedia Computer Science*, 184, 148–155. https://doi.org/10.1016/j.procs.2021.03.026

Névéol, A., Dogan, R., & lu Zh. (2010). Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction. *Journal of biomedical informatics*, 44, 310–318. https://doi.org/10.1016/j.jbi.2010.11.001

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., & Roberts, I. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatic*s, 42. https://doi.org/10.1016/j.jbi.2008.12.013

Scribendi. (May, 2023). *The rules for writing numbers in English*. https://www.scribendi.com/academy/articles/when_to_spell_out_numbers_in_writing.en.html?session_token=eyJ0aW1lIjoxNjY4NDg0Mjk4MTQ3LCJob3N0Ijoid3d3LnNjcmliZW5kaS5jb20iLCJyZWZlcmVyIjoiaHR0cHM6Ly93d3cuc2NyaWJlbmRpLmNvbS9hY2FkZW15L2FydGljbGVzL3doZW5fdG9fc3BlbGxfb3V0X251bWJlcnNfaW5fd3JpdGluZy5lbi5odG1sIn0%3D

Shrestha, I., & Dhakal, Sh. (2021). Fine-grained part-of-speech tagging in Nepali text. *Procedia Computer Science,* 189, 300–311. https://doi.org/10.1016/j.procs.2021.05.099

Soumia, B., Cherroun, H., & Abdelali, A. (2017). Altruistic crowdsourcing for Arabic speech corpus annotation. *Procedia Computer Science*, 117, 137–144. https://doi.org/10.1016/j.procs.2017.10.102

UCREL NLP GROUP. (May, 2023). CLAWS. University of Lancaster, UK. https://ucrel.lancs.ac.uk/claws/

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University – Computer and Information Sciences*, 34 (4), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

Yamamoto, K., Kudo, T., Konagaya, A., & Matsumoto, Yu. (2005). Use of morphological analysis in protein name recognition, *Journal of biomedical informatics*, 37, 471–82. https://doi.org/10.1016/j.jbi.2004.08.001

Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia – Social and Behavioral Sciences,* 95, 20–32. https://doi.org/10.1016/j.sbspro.2013.10.618

Статья поступила в редакцию 10.06.2023; одобрена после рецензирования 27.11.2023; принята к публикации 04.12.2023.
The article was submitted 10.06.2023; approved after reviewing 27.11.2023; accepted for publication 04.12.2023.