Piotr Pęzik

PROVIDING CORPUS FEEDBACK FOR TRANSLATORS WITH THE PELCRA SEARCH ENGINE FOR NKJP

Abstract: The present paper introduces the PELCRA search engine for the National Corpus of Polish (PSEN). This highly scalable corpus search tool provides access to concordance and collocation search results in a variety of output formats, including downloadable spreadsheets, compressed URL-s, integrated browser plug-ins and web services. Apart from outlining the general functionality of this tool, I demonstrate the usefulness of PSEN slop factor queries in verifying the phraseology of translated texts. This specific example is meant to illustrate the general possibility of ensuring easy and universal access to corpus evidence through the combination of a robust query syntax and compressed Uniform Resource Locators (URLs). The search engine is freely available at *www.nkjp.uni.lodz.pl*.

Keywords:

1. National Corpus of Polish

The National Corpus of Polish (NKJP) is a research project funded by the Polish Ministry of Science and Higher Education (project number: R17 003 03). Its main objective is the construction of a 300 million word (MW) balanced corpus of modern Polish. The total pool of texts used to sample the balanced corpus from will reach 1 billion words by the end of 2010. More detailed information on the recent developments in the NKJP can be found in (Przepiórkowski *et al*. 2009) and on the project's web site at *www.nkjp.pl*.

The sheer size of the acquired data as well as the rich Polish morphology pose special challenges for the retrieval engines used in the project. The PELCRA NKJP corpus search engine (PSEN) is one of the two tools used for the

NKJP[1]. Based on a combination of Apache Lucene[2] and relational database technologies, it scales well with the size of the corpus; an uncached query for *raczej* matching 85 429 paragraphs in a 500 MW version of the corpus retrieves the first one hundred results in 0.063 seconds. This sort of performance can currently only be achieved in index-based retrieval systems.

2. Exporting and sharing results

For the sake of clarity, it makes sense to introduce a feature of PSEN used throughout the paper to provide a permanent reference to the original result screens obtained for submitted queries. The results of all search operations supported by the engine can be saved and shared in the form of compressed URLs. Having typed in a complex query, the user can click the *URL* button (shown in Fig. 1) to generate a compressed reference to the result set. As an example, the query for the noun *prawda*, with additional sorting and grouping options is compressed to the following form:

http://nkjp.uni.lodz.pl/?q=ygtpdvf

Such a compressed link can be easily shared or saved for future reference. Typing in the compressed link in a web browser's address bar should bring up precisely the original result screen together with the completed search form. The reader may find the discussion of the PSEN query syntax easier to follow by typing in the compressed URLs listed next to each of the examples I introduce below. This feature of PSEN also eliminates the need for including space-consuming screenshots. The use of URLs as references to concordance results was previously introduced by (Gaskell & Cobb 2004) and also implemented in the *IFAConc* tool.

[1] *Poliqarp for NKJP* is another search tool available for the NKJP data. It supports more complex types of linguistic search at various levels of annotation (http://chopin.ipipan.waw.pl/poliqarp/).

[2] http://lucene.apache.org

3. Basic PSEN Query syntax

Given a basic introduction, the PSEN query syntax is relatively easy to learn. Single term queries such as:

rewolucja (http://nkjp.uni.lodz.pl/?q=yz3nckf)

can be expanded both orthographically:

rewolucj* (http://nkjp.uni.lodz.pl/?q=yc4y7r8)

and morphologically, as in:

rewolucja** (http://nkjp.uni.lodz.pl/?q=yjckk4g).

The morphological expansion of a query term saves a lot of typing in languages such as Polish, where a single verb may have dozens of inflectional variants.

Multi-term queries for phrases can be specified by simply typing a sequence of query terms. It is important to remember that for a term to be morphologically expanded, it needs to be entered as a lemma, even if a phrase query is not intended to retrieve the required base form itself. To illustrate, the query for *Wielka Brytania* (Great Britain), i.e.:

Wielki** Brytania** (http://nkjp.uni.lodz.pl/?q=yz2r3yo)

is clearly meant to match the occurrences of the feminine forms of the adjective *wielka* even though the base masculine base form *wielki* is required by the query syntax.

The pipe symbol | can be used to specify morphological, orthographic or lexical variants of a term. For example, to search for all inflections of the noun *łza* (tear) and its diminutive variant *łezka*, one could formulate the following query:

łza**|łezka** (http://nkjp.uni.lodz.pl/?q=ycvghd2).

The orthographic expansion wildcard can also be prefixed to the search term. For instance, to check the productivity of the *-essa* suffix, which is often used in Polish to create feminine noun variants, the following query could be used:

*essa** (http://nkjp.uni.lodz.pl/?q=ygvfhn5)

resulting in matches such as *stewardessa*, *poetessa* or *hostessa*.

4. Slop factor and term order

The basic query syntax is complemented by a number of options available in the search form. The following screenshot shows the options currently available in the basic search form:

Fig. 1. Basic options search form

The first two options, i.e. *Slop* and *Preserve order* can be used to specify the maximum slop factor and the word order restriction respectively. The slop factor indicates how many words may occur between the query terms. Increasing the slop factor generally increases the recall of the result sets when searching for flexible collocations, at the expense of their precision. As an example, the query shown in the screenshot is intended to find collocations of the nouns *łza* (tear) and *oko* (eye) with the verb *zakręcić*, as in "*łza zakręciła się w jej oku*". The phrase would roughly translate into English as "***a tear welled up** in her **eye***".[3]

To further increase the recall of the query, the *Preserve order* box can be unchecked to indicate that the order in which the query terms are entered in the form does not have to be preserved in the matching results. Quite conveniently, a single query such as:

*łza**|*łezka** oko** kręcić**|zakręcić***

[3] For an example of an equivalent query against the British National Corpus, see http://nkjp.uni.lodz.pl/?q=ya7x6u6.

where the terms *łza*** can be labelled as *1A*, *łezka*** as *1B, oko*** as *2A*, *kręcić*** as *3A*, and *zakręcić*** as *3B*, can thus result in a set of permutations of this collocation such as:

kręcą się łzy w oczach (3A 1A 2A)
kręcą się w oczach łzy (3A 2A 1A)
łezka kręci się w oku (1A 3A 2A)
łezka się w oku kręci (1B 2A 3A)
oczach jej zakręciły się łzy (2A 3B 1A)
zakręciła się łezka w oku (3B 1B 2A)

(see: http://nkjp.uni.lodz.pl/?q=p33wg5).

Other options in the advanced search screen include 2-level sorting and grouping of results. The latter option is particularly useful when a large result set is retrieved; a maximum number of results from each register, text or date can be specified, thus providing an outline of the distribution of the matching words and phrases across registers and time periods. For example, limiting the maximum number of occurrences of the verb *ściemniać* (whose literal meaning roughly corresponds to the English verb *to dim*) to 3 per year (http://nkjp.uni.lodz.pl/?q=yzoezf8) shows that one of the first occurrences of *ściemniać* in the underlying corpus in its relatively recent metaphorical sense of "obfuscating" can be found in a text published in 1996.

It is also possible to search with metadata criteria, including genre, medium, text title and publication date. Particularly useful may be the option of limiting the result set to paragraphs matching a Boolean query specifying contextual keywords. For example, it is possible to search for all the occurrences of the word *połączenie* (connection) as long as it occurs in a paragraph containing at least one of a set of obligatory keywords, such as *kolej* (railway), *pociąg* (train), *autobus* (bus), but none of a set of disqualifying keywords such as *telefon, sieć* (network) or *serwer*:

http://nkjp.uni.lodz.pl/?q=ykm43pv

This option enables basic word sense disambiguation, which comes in handy whenever the user is only interested in one sense of a polysemous word.

5. Collocation extraction module

A separate module of PSEN facilitates the extraction of collocations. In order to retrieve a set of potential collocations, it is necessary to specify the node terms using the query syntax introduced above. As an example, the following definition of a node term:

*bezczelny*** (http://nkjp.uni.lodz.pl/?q=qfrs93)

will be expanded into the set of all inflections of the adjective *bezczelny*. Interestingly, the node of a collocation can be defined as a sequence of terms, thus giving some possibility of extracting multiword collocations, e.g.:

*wymigać** *się od* (http://nkjp.uni.lodz.pl/?q=pe2q27).

When defining a single term collocate of a given node, it is possible to define its basic part of speech category as well as some positional restrictions, such as the distance from the node. To extract a set of potential adjectival collocations of the noun *zdrada*, the following query could be run:

zdrada** (http://nkjp.uni.lodz.pl/?q=onbdzr)

with the part of speech option set to *adjective*.

6. Application of PSEN in translation teaching

Corpus search software is often claimed to play an important role in a translator's workshop. This claim is based on the generally recognized applicability of corpus tools in the study of lexis and the fact that the job of a translator involves making quick and accurate lexical and phraseological choices. This potential capability of corpus tools is particularly important for beginning translation students, who often find it difficult to break free from the shackles of formal equivalence and develop dynamic equivalence skills at the phraseological level.

Although intuitively, this application of corpora seems reasonable and many translators are indeed frequent corpus users, it is not always obvious how beginning translation students can be encouraged to fully benefit from

representative language resources. A general introduction to corpus linguistics is often not enough for translation students to develop a sustained reliance on corpus evidence. Students often lose their interest in corpus resources for a number of reasons. First of all, setting up a desktop corpus application for a large underlying corpus can be a nuisance. Depending on the speed of the computer used and the implementation of the corpus software, it may also be difficult for the software to compete with *Google* in terms of providing instant and universal access to simple examples of word or phrase usage, which is probably the most frequent reason why a translator would refer to a corpus. Moreover, the query syntax of generic corpus tools may offer little advantage over internet search engines when it comes to phraseological queries. Finally, the hassle of reproducing more complex corpus queries and result sets may also discourage both translation students and teachers from a frequent use of corpora. In the remaining part of the paper I address these issues by focusing on two features of PSEN which facilitate the use of corpus evidence in a translation course curriculum.

7. Slop factor phraseological queries

Most translation students are perfectly willing to admit a logical mistake or obvious problems with single-word lexeme equivalence in their assignments. It is usually more difficult to convince them about phraseological incongruities in the target text, and especially so when they are translating into their mother tongue. As native speakers they do not feel less entitled to phraseological judgments about their mother tongue than their teacher.

Although phraseological restrictions are often purely idiomatic, they can equally well be positioned on the grammatical end of the continuum of lexico-grammar (Sinclair 1991). Therefore the difficulty of identifying and explaining phraseological errors in target texts stems not only from purely idiomatic restrictions but also from lexico-grammatical intricacies. Without instant access to a convincing body of evidence, discussing either of the two types of incongruities is often perceived as an exercise in semantic hair-splitting. The advantage of using PSEN in such a situation is that it supports high recall phraseological queries which can be compressed to short URLs, which in turn can be used to annotate a suspicious lexical or phraseological structure.[4] Let us

[4] The use of URL-based concordance references to provide feedback for writing errors was previously discussed by (Cob & Gaskell 2004).

consider the following excerpt from a BBC news article on the recent celebrations of the demise of the Berlin Wall with a corresponding translation provided by one of my translation students:

> ORIGINAL: "History is palpable and alive here. The peaceful revolution of the fall of the Wall 20 years ago ***paved the way*** to an unprecedented transformation of Berlin," Mayor Klaus Wowereit said.
>
> TRANSLATION: *"Tutaj historia jest wyczuwalna i ciągle żywa. Pokojowa rewolucja związana z upadkiem muru dwadzieścia lat temu niejako* ***wybrukowała drogę do*** *bezprecedensowego procesu przemiany Berlina,' powiedział burmistrz miasta Klaus Wowereit."*

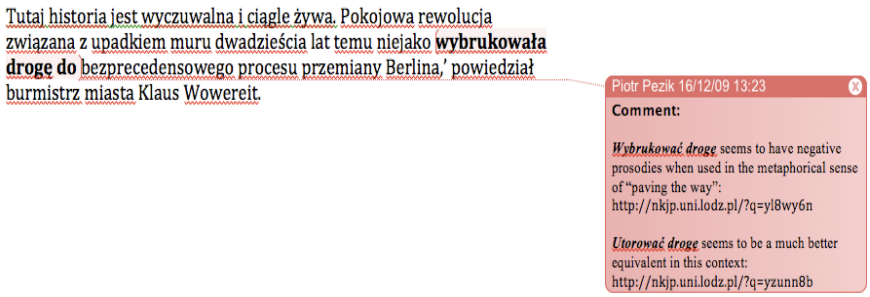
One problematic phraseological item in this translation is the expression *wybrukować drogę*, which is a fairly literal translation of the phrase *to pave the way* used in the original.[5] Most of the native speakers of Polish presented with this translation have found the phrase in question "artificial", but none have attempted a more precise explanation without corpus evidence. One reason why a more definitive explanation is difficult to provide in this case is the fact that the verb + noun collocation *wybrukować drogę* and its variants do occur in Polish[6], and thus it cannot be dismissed as a totally artificial equivalent. A closer look at the examples retrieved from the NKJP data reveals, however, that it is frequently used in the literal sense of "paving a road" and when used metaphorically, it tends to have rather negative prosodies as in *droga do piekła wybrukowana jest dobrymi intencjami* (*the road to hell is paved with good intentions*). Furthermore, another phraseological equivalent *utorować drogę* seems to fit much better in this context[7]. Although based on a slightly different metaphor, it is used almost exclusively in the figurative sense of enabling or facilitating actions and achievements.

Using PSEN it is possible to reference both positive and negative corpus evidence with two compressed URLs. As shown in the screenshot below, the teacher marking up a translation assignment can thus make evidence-based and verifiable comments about the student's phraseological choices:

[5] Incidentally, in contrast to English, Polish does not distinguish between *droga* as a *way* and *droga* as a *road*.

[6] http://nkjp.uni.lodz.pl/?q=yl8wy6n

[7] http://nkjp.uni.lodz.pl/?q=yzunn8b



All the students need to do to access the relevant corpus evidence is type the URL provided in their browsers' address bars and hit the enter key. Obviously, a similar sample could be obtained with a few queries sent to a standard concordancer. However, the simplification of tedious search procedures by means of compressed links leaves students with no excuse for ignoring the corpus evidence available.

The collocation extraction module of PSEN can be used in a similar fashion to provide examples of lexico-grammatical restrictions which are difficult to retrieve with a simple concordancer query.

Let us consider an example news article about a space mission aimed at studying the Earth's water cycle. A student's translation of this passage into Polish uses a relative clause based on the verb phrase *mieć na celu*, which can be literally translated as *to have an objective*. Although the choice of the phrase seems to fulfil formal equivalence criteria, it raises difficult to explicate lexico-grammatical suspicions. The problem with *mieć na celu* in this context can be traced to the use of *statek kosmiczny* (spacecraft) as its subject:

> ORIGINAL: *The SMOS spacecraft launched on Monday* ***to study the Earth's water cycle*** *has passed a key mission milestone.*
>
> TRANSLATION: *Wystrzelony w poniedziałek* ***statek kosmiczny SMOS mający na celu*** *obserwację obiegu wody na Ziemi osiągnął już najważniejszy etap swej misji.*

Somehow, *statek kosmiczny* does not sound right as the subject of *mieć na celu*, but this incongruity can only be fully exposed by examining the subcategorization frames of the idiomatic verb phrase in question. The collocation extraction module of PSEN makes it particularly easy to obtain a list

of nouns preceding a word or phrase, have them sorted by statistical significance and saved as a compressed URL:

http://nkjp.uni.lodz.pl/?q=yhamox2

A brief look at the results shows that the great majority of nominal collocates preceding the phrase *mieć na celu* are hyponyms of the notions of *action* or *effort.* There seems to be a semantic restriction on the range of subjects of this lexical item, which would be difficult to identify with a simple concordancer, due to the lexico-grammatical rather than purely lexical nature of the restriction (Biber et al. 2008). In other words, the fact that *statek kosmiczny* does not occur in the NKJP corpus as a subject of *mieć na celu* does not convincingly prove that such a combination is impossible. On the other hand, generating a list of typical subjects of the phrase revealing its strong preference to describe the intended purpose of abstract actions rather than physical objects does seem to confirm the seemingly inexplicable intuition signalled by most of the speakers of Polish presented with this translation.

Again, the advantage of using PSEN to bring up the problem is not only the collocation extraction module, but also the possibility of generating direct links to support the teacher's comment with a wealth of corpus evidence. Incidentally, should the teacher misinterpret the corpus evidence available, such a marking procedure may turn out to be a double-edged weapon. An evidence-based discussion of translation choices, however, would be a rather desired development in a university class.

8. Summary

The PSEN syntax supports slop factor and relaxed word order functions, which increase the recall of phraseological queries. In other words, a single query can be formulated to fetch a sample of occurrences of a phrase regardless of the order and position in which its individual constituents and their lexical, orthographic or grammatical variants occur. As illustrated above, such evidence can prove to be indispensable in identifying both lexical and lexico-grammatical incongruities. Together with the collocation extraction module, this functionality can be of real interest to translators, linguists and lexicographers.

The simplification of online search tools sometimes referred to as *googlization* can be defined as an interface design strategy following the assumption that search engine users only need to learn complex syntax and

advanced options when absolutely necessary. The familiarity of web-based interfaces also means and that users tend to choose fast and universal browser-based access interfaces over less familiar desktop clients. PSEN follows this philosophy not only by hiding away the advanced options from casual users, but also by providing a way to encode complex searches in compressed URLs, which can be used to reference and share corpus findings.

The present paper shows how the combination of slop factor queries and URL compression can be used to retrieve and share corpus evidence in a way that leaves virtually no excuse for ignoring it. This solution has one application in translation teaching, but it is not difficult to imagine similar applications in other standard use cases of corpus linguistics techniques, such as dictionary making or language teaching. The PELCRA Search Engine for NKJP is hopefully an example of how state-of-art technologies can be used to provide universal access to corpus evidence.

REFERENCES

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Biber, D., Conrad, S., and R. Reppen (2008). *Corpus Linguistics: Investigating Language Structure and Use. Cambridge Approaches to Linguistics.*

Gaskell, D. and T. Cobb (2004). *Can learners use concordance feedback for writing errors?* http://www.lextutor.ca/cv/conc_fb.htm.

Przepiórkowski, A., Górski, R., Łaziński, M. and P. Pęzik (2009). *Recent Developments in the National Corpus of Polish.* In: J. Levická and R. Garabík (eds.). *NLP, Corpus Linguistics, Corpus Based Grammar Research: Proceedings of the Fifth International Conference, Smolenice, Slovakia, 25–27 November 2009*. Slovko 2009: 302–309.