Elżbieta Mańczak-Wohlfeld   https://orcid.org/0000-0002-7839-4957

Jagiellonian University

Alicja Witalisz   https://orcid.org/0000-0002-2256-1269

Pedagogical University of Krakow

Anglicisms in The National Corpus of Polish: Assets and Limitations of Corpus Tools

Abstract

While electronic corpora may not seem adequate sources for anglicisms retrieval, since despite promising attempts they still lack readily available and efficient tools for foreign loans identification, they are indispensable in a systematic verification of the use of preidentified loans. The article offers an assessment of an electronic corpus of Polish in reference to its usefulness for the study of English loans. Though we test a selected corpus and its tools, and use Polish anglicisms as exemplifications, the findings presented in the article pertain to other large corpora and anglicisms in other languages. Corpus tools allow for a multi-dimensional analysis of loans, yet they fail to meet the requirements of more in-depth analyses of anglicisms, related to their semantics and structure. The limitations of corpora tools will be illustrated with authentic attempted-but-failed corpus searches.

Keywords

anglicism, National Corpus of Polish, English borrowing, loanword adaptation

Streszczenie

Pomimo obiecujących badań automatyczna ekstrakcja anglicyzmów z wykorzystaniem narzędzi dostępnych w elektronicznych korpusach językowych wciąż nie jest możliwa. Mimo to wyszukiwarki korpusowe są nieodzownym narzędziem w systematycznej weryfikacji użycia anglicyzmów wyłuskanych metodą tradycyjną. W artykule omówiono zarówno funkcjonalność, jak i niedoskonałość narzędzi dostępnych w Narodowym Korpusie Języka Polskiego w odniesieniu do badania anglicyzmów różnych typów oraz ich z góry zdefiniowanych cech. Niedostatki narzędzi, związane głównie z semantyką zapożyczeń, zostały zilustrowane konkretnymi przykładami anglicyzmów.

Słowa kluczowe

anglicyzm, Narodowy Korpus Języka Polskiego, zapożyczenie angielskie, adaptacja zapożyczeń

1. Introduction

Research on anglicisms prevalent in European languages has continually pre-occupied language contact linguists who until recently have been (and still are) engaged in a painstaking manual search for English loans in the printed press, TV and radio programmes, public speeches, informal conversations, youth's slang etc. With the advent of electronic corpora and data processing tools, this work has become less arduous and more efficient. Yet the very identification of a loan, which is the starting point in the analysis of the anglicism adaptation and usage in any recipient language, still remains the researcher's responsibility and depends on their knowledge-based human skills that have not, as yet, been successfully copied and replaced with artificial intelligence. Efforts are being made to develop data processing tools that would automatise the process of anglicism identification and extraction. Promising attempts at the automatic retrieval of lexical anglicisms in some Romance and Germanic languages have been made, exploiting manually extracted bi- and trigrams that are most typical of English (Furiassi and Hofland 2007; Furiassi 2008). Generally, the data processing tools for the automatic extraction of English loans take advantage of the differences in orthography between the languages in contact and use grapheme typicality algorithms, as well as dictionary-based methods and word-formation regularity (Andersen 2005, 2011, 2012). A complementary approach to automatic anglicism retrieval in Norwegian uses machine-learning methodology and a data-driven frequency approach (Losnegaard and Lyse 2012). The attempts have been (partially) successful for the most frequent anglicism type, i.e. for unadapted lexical loans whose formal foreignness makes their identification easier also in a manual extraction. Much more demanding, if any such attempts are undertaken, will be the development of data processing tools for an automatic identification and extraction of semantic loans and loan translations, formally covert by the native lexical material they are composed of.

Therefore despite encouraging experiments, anglicisms still have to be manually predefined by the researcher, irrespective of the corpus size (Davies p.c.).[1] Once they are predefined, corpus tools offer a wide range of possibilities for the automatic corpus-based and corpus-driven analysis of loans and the retrieving of their various types and variants. An effective study of anglicisms and an accurate assessment of the degree to which English has penetrated the recipient languages requires the supplementing of the traditional excerption of loans and the typically quantitative analyses with a systematic

---

[1] Mark Davies, creator of the 14-billion-word iWeb: The Intelligent Web-based Corpus; personal communication at *X International Conference on Corpus Linguistics*, Cáceres, Spain, 9–11 May 2018.

corpus-assisted study of the various and often unexpected outcomes of language contact, which is not possible without the assistance of large, diversified and continuously updated electronic corpora and corpus tools.

In a recent publication, Lewandowska-Tomaszczyk and Wilson (2018: 179) argue that "when it comes to semantic and pragmatic annotations of meanings in use, particularly in large corpora, adequate corpus tools are not yet fully developed." While corpus-assisted research is revealing as for the use and adaptation of anglicisms in the recipient language, as will be illustrated in Section 3, we attempt to identify those areas of foreign loan research in which the available corpus tools turn less efficient or quite couterproductive.

We first offer a brief historical overview of the work on electronic databases of contemporary Polish. Having selected one of the available corpora of Polish and provided arguments for its selection, we proceed to a discussion and exemplification of the strengths and limitations of its tools in the study of Polish anglicisms. In the final sections, we draw conclusions and define areas of corpus-assisted anglicism research that pose problems for language contact researchers, related to automatic loan detection and corpus-assisted research on foreign loans, thus making a plea to corpus linguists and IT experts for data processing tools capable of (even more) effective loan retrieval.

2. Corpora of Polish – a historical overview

The present section summarises the attempts that have been undertaken in the last half a century to design and implement a comprehensive and well-balanced corpus of contemporary Polish, to ultimately point to the corpus that offers the most advanced tools for foreign loans analysis, despite the gap in its updating and a handful of deficiencies.

In 2001, in the Introduction to *A Dictionary of Anglicisms...*, Görlach states that:

> The comparative method and the time schedule have also precluded basing our statements (including those on currency) on text corpora. There are doubts about the representativeness of corpora ... and the methodological problems proliferate with any cross-linguistic analysis. Moreover, for many languages here included such corpora would have had to have been put together from scratch – so there was really no choice but to base statements about style and currency values on the introspection of the collaborators and their informants, combined with data in recent dictionaries. (Görlach 2001: XVI)

Such an approach to the study of foreign loans was certainly right at the time, as some languages, including Polish, did not have any reliable corpora when Görlach's lexicon of anglicisms was being compiled. The first dictionary of

frequency (Kurcz et al. 1990) came out in 1990 and was based on a corpus in the modern sense of the word, namely on frequency lists (Kurcz et al. 1974–1977). The data, coming from texts published in 1963–1967 and divided into five sections: essays, news, scientific texts, fiction, and plays, contained bibliographical descriptions of the sources. Each word was tagged with its base form and selected morphological properties; sentence boundaries were marked (clip.ipi.pan.waw.pl/PL196x). The frequency dictionary altogether contained around half a million running words, which were somewhat outdated. In the range list of absolute frequency out of 10,355 enumerated lexical items only 59 constituted English loans (40 loans and 19 derivatives) (Mańczak-Wohlfeld 2004).

The first corpus of Polish accessible to the public appeared in the early 2000s and was authored by the researchers from the Institute of Computer Science of the Polish Academy of Sciences (IPI PAN). The IPI PAN corpus was a collection of over 100 million running words from morphosyntactically annotated texts and contained a balanced subcorpus (Przepiórkowski 2004: 5). It followed current standards and best practices in corpus linguistics, and could be approached via Poliqarp search engine. As Przepiórkowski, the coordinator of the project, notes:

> The current version of both the corpus and the tools is called here a *preliminary version:* we are painfully aware of various inadequacies of the corpus and the tool... Taking into consideration the sheer size of the corpus, and the limited resources at the disposal of the project, it was impossible to verify the results of the automatic conversion of the incoming texts into the XML format, or the results of morphosyntactic, structural and metadata annotation... The IPI PAN corpus in its current form is a typical opportunistic corpus, containing various genres in unbalanced proportions. (Przepiórkowski 2004: 6–7)

Previous corpus research was conducted in the 1990s in various academic centres. The oldest compilation of a Polish corpus, launched by Barbara Lewandowska-Tomaszczyk from the University of Łódź in cooperation with Tony McEnery from Lancaster University, goes back to 1995. The researchers worked at the time when the then largest corpus of English, the British National Corpus, was being created. This corpus of Polish, referred to as the PELCRA (Polish and English Corpora for Research and Application), contains 100 million words, all of which are available to the public (http://korpus.ia.uni.lodz.pl or http://pelcra.pl).

In 1997, the dictionary section of the PWN publishing house in Warsaw started to work on a new corpus of Polish, known as the PWN corpus (korpus.pwn.pl). It consists of raw material, i.e. linguistically unannotated texts, and contains 100 million words of which 40 million are available at http://korpus.pwn.pl. The PWN corpus aided the work on, among others, two

general-purpose lexicons of Polish, by Bańko (2000) and by Dubisz (2003). The PWN corpus was intended to form the basis for lexicographic descriptions as well as be a source of varied examples. The texts included in the corpus are diversified thematically and stylistically as well as representatively from the point of view of the Polish literary tradition, which accounts, among others, for the inclusion of all the school reading lists.

The third centre to continue work on the corpus of Polish was the Institute of the Polish Language of the Polish Academy of Sciences in Kraków, where scholars headed by Rafał Górski developed an internal corpus available only for research carried out at the Institute (www.ijp-pan.krakow.pl). Originally the corpus was meant to be used as the basis for a large lexicon of Polish.

In conclusion, none of the corpora presented above was large enough, diversified enough, representative enough or, at the same time, morphosyntactically diversified to constitute a comprehensive and well-balanced corpus of Polish.

In 2006, the Linguistic Committee at the Polish Academy of Sciences initiated a consortium that would work on the National Corpus of Polish. The project was launched in collaboration with the Institute of Computer Science, PAN, Institute of Polish, PAN, the PWN Publishing House and the Chair of Computational and Corpus Linguistics at the University of Łódź, and coordinated by Adam Przepiórkowski from IPI PAN. The National Corpus of Polish (Narodowy Korpus Języka Polskiego, NKJP) was ready in 2011. It contains 1.5 billion words taken from heterogeneous sources, including daily and specialised press, literary works, non-fiction, spoken and electronic texts. The thematic and genre diversity in the spoken texts coincides with the balance of genders, ages and regions. The earliest texts are dated for the early 20th century, but 80% of the materials are post-1990. It offers two search engines: Poliqarp & PELCRA, that have both shared and distinctive tools (Pędzik 2012; Przepiórkowski et al. 2012, 2017).

The most up-to-date corpus of contemporary Polish was opened to the public in 2016 (monco.frazeo.pl). Designed by the Department of Computational and Corpus Linguistics at the University of Łódź, MoncoPL is a 5-billion-word corpus that is daily updated, drawing from a thousand Polish webpages, including electronic versions of newspapers, magazines, TV and radio programmes, as well as popular portals and private blogs.

For the usefulness tests carried out in the subsequent sections of the article, out of the two largest and most up-to-date collections of contemporary Polish we will employ the National Corpus of Polish (NKJP) in its full 1.5-billion-word version, for its size, thematic and genre diversity, but most of all for its advanced search tools that will be exploited for the advanced analyses of anglicisms.

3. Strengths of corpus tools in the analysis of foreign loans

Any analysis of loan properties, adaptation and usage, aided by electronic data processing tools, must be preceded with the very identification of an anglicism in the recipient language. Attempts at automatic loanword identification are mentioned in the Introduction and further described in Section 4. Once an anglicism has been identified with the manual methods of excerption, we begin with determining the properties of anglicisms that are verifiable with the use of the two complementary NKJP search engines, PELCRA and Poliqarp. At the launch of the search the assumption is that the study of loans will be largely corpus-based, i.e. the corpus will be exploited to test knowledge- and research experience-based hypotheses about anglicism types, features, and adaptation.

3.1. Loan and loan type verification

With a list of predefined anglicisms at hand, the NKJP offers tools that allow for the verification of the use and institutionalisation of various types of anglicisms, yet with restrictions on some loan types. While searching for concordances displaying English loanwords, i.e. English lexemes borrowed with both form and meaning, such as e.g. *jazz*, *coming out*, *VIP* and *fake news*, does not pose a problem, an attempt to identify semantic loans that are foreign senses borrowed from English polysemous lexemes, is a challenge for the researcher, who engages in a time-consuming process of separating contexts in which a native word is used in a new, foreign sense (see Section 4.4). The verification of loan translations, i.e. direct more or less exact word-for-word translations of foreign multi-word expressions, such as e.g. Pol. *miękka władza* (< Eng. *soft power*), proves unproblematic, unless the loan translated idiomatic expression coincides with a native loose syntactic phrase that is homonymous to the former, e.g. Pol. *gorący ziemniak* (< Eng. *hot potato*), *szklany sufit* (< Eng. *glass ceiling*) (see Section 4.3).

In the middle of the continuum between overt loanwords and covert loan translations and semantic loans lie loanblends that are half-translations of foreign polymorphemic expressions, e.g. Pol. *długi drink* (< Eng. *long drink*), Pol. *anioł biznesu* (< Eng. *business angel*). Corpus tools allow for the verification of preidentified loanblends, and also for the automatic extraction of other loanblends that make use of the same foreign element, provided the two-morpheme expression is spelt as one word, which allows a query involving predefined loanblend constituents, e.g. *-holik* as in Pol. *pracoholik* (< Eng. *workaholic*), Pol. *zakupoholik* (< Eng. *shopaholic*).

In the attempts at automatic identification of anglicisms (Andersen 2012: 126), hybrid compounds (composed of English-Norwegian lexical material)

are not identified as anglicisms due to the recipient language elements having native characteristics. Therefore semi-automatic extraction of hybrid compounds gives better results, provided we are vested with a set of preidentified or potential English-originating compound constituents. The same semi-automatic method for hybrid loan extraction may be successfully used for the excerption of multi-morphemic expressions derived in the recipient language with English morphological loans, such as borrowed affixes, e.g. *-ing*, and combining forms, e.g. *cyber-*, *e-*, *-gate*, provided the researcher is aware of the foreign bound morpheme productivity in the recipient language.

Corpus tools allow not only for the verification but also for the excerption of instances of one-word hybrid creations (with the foreign element predefined) that have been stimulated by language contact, e.g. Pol. *szafing* (Pol. *szafa* 'wardrabe' + Eng. *-ing*), Pol. *ciucholand* (Pol. *ciuch* 'clothing' + Eng. *-land*), *randkoholik* (Pol. *randka* '(romantic) date' + Eng. *-holic*). These are often *hapax legomena*, difficult to detect in a manual search, and have no discoverable models in the source language.

Using the same processing tool lets us find more instances of compound loanwords that share a lexical element, as in the case of Pol. *biznes* (< Eng. *business*) that reappears in e.g. Pol. *bizneswomen* (< Eng. *businesswoman*), Pol. *show-biznes* (< Eng. *show business*), Pol. *biznes partner* (< Eng. *business partner*), Pol. *biznes lunch* (< Engl. *business lunch*), Pol. *biznes class* (< Eng. *business class*), though in the case of compound loanwords spelt separately it is a rather daunting task.

Despite its advanced tools, to be described in the following sections, the NKJP, having been last updated in 2012, lists neither the newest high-frequency anglicisms, e.g. Pol. *selfie*, *hasztag/hashtag*, nor some well-assimilated English loans used in the media, e.g. Pol. *crowdfunding*, *followersi* (< Eng. *followers*), *snapchatowanie* (< Eng. *snapchatting*). This is where we may profit from the MoncoPL corpus, updated daily with web-driven data.

3.2. Loan frequency measurement

Regardless of the doubts concerning the notion of word frequency and the representativeness of corpus-derived statistics (cf. Moon 1998: 7; Schmid 2010: 125–126), corpus tools test both absolute and relative word frequency. The PELCRA search engine determines the frequency for the preidentified anglicisms of various types in two separate searches, either including or excluding loan inflectional forms and loan-based derivatives. The assets of the NKJP tools will be illustrated with the lexeme *leasing*, which is a well-established English loanword in Polish, unadapted graphically and thus recognisable as a foreign word, partially adapted phonologically through the substitution of English vowels with their closest Polish equivalents, and the vocing

and devoicing of the middle and final consonants, respectively. It is morphologically integrated in Polish having a full inflectional paradigm in the singular and serving as a base for native adjectival, verbal and nominal derivatives.

The frequency of *leasing* in Polish according to the NKJP is 3,640, if the search is limited to the basic, Nominative form. In highly inflectional languages such as Polish, determining the frequency of a loan without a tool allowing for finding its inflected forms would be deceptive. While both NKJP tools effectively display concordances for the inflected forms of a loan (pertaining to case and number), PELCRA provides a more readily available frequency count (5,864 occurrences). If the frequency search is extended to include loan-based derivatives, the frequency of *leasing* in Polish rises to 7,868. Yet automatic determining of the frequency of just loan-based derivatives (without inflected forms) is not possible.[2]

NKJP tools are able to determine loan frequency in respect of time, register type included in the corpus (e.g. formal, literary, conversational) and source type (e.g. press, literature, Internet, spoken language), which proves particularly useful for language contact scholars who examine loans from the sociolinguistic perspective. An automatic composition of a loan diachronic profile with the frequency tool points to the time period in which the anglicism might have been borrowed and its frequency was highest. The first, single at the time, corpus attestation of *leasing* in its base form in Polish goes back to 1988, whereas the peak frequency was noted in 2000 and 2001 with raw occurrences reaching 383 and 428, respectively, and in 1995 if measured for every 1,000 paragraphs (0.253). The frequency measurement tool is not entirely deficiency-free. In the case of covert loans, separate concordances can be obtained neither for the semantic loans nor for loan translations that are homonymous to the recipient language loose syntactic phrases (see 4.3 & 4.4).

3.3. Verification of loan graphic variants

The frequency of loan use rises if we take into account its graphic variants (or misspellings), yet the researcher has to predefine the alternative (or self-invent the potential) spelling of a loan. For the loan example we use in this study, the most obvious Polish graphic variants of *leasing* include *lizing* (reflecting its Polish pronunciation) with 35 hits and *lising* with 6 hits, increased to 52 and 17, respectively, if the search includes inflected and derived forms.

A fully automatic search for the alternative graphic variants of *leasing* with the NKJP tools using the symbols ? and . (dot) in Queries [li?ing] in PELCRA and [li.ing] in Poliqarp, brings dubious results, as it displays only 2 concordances for *lising* out of 6 found in a predefined search.

---

[2] We would like to thank the anonymous reviewer for valuable prompts regarding the NKJP tools.

3.4. Verification of loan morphological adaptation

Empirical testing of the morphological adaptation of English loans, which may embrace, among others, native suffix addition, suffix replacement, loan inflectionality (in the case of inflectional recipient languages), and the formation of loan-based derivatives, is verifiable with NKJP search engines that are complementary. Both PELCRA and Poliqarp effectively find concordances for morphologically adapted loans, separating inflected forms of loans from loan-based derivatives. Poliqarp additionally provides tagging for part of speech, number, case and gender (cf. e.g.: **leasingiem** [leasing:subst:sg:inst:m3]). Both tools allow for finding loan-based derivatives through Queries [leasing*]/[*leasing*] and [!base=leasing & base=.*leasing.*] that display *leasing*-based derivatives (incl. their inflected forms), which are mostly cases of affixation (e.g. Pol. *leasingowy* (adj.), Pol. *poleasingowy* (adj.) 'post-leasing'; Pol. *leasingować* (v.) 'to lease'; Pol. *wyleasingował* (v., 3rd p.sg., past tense, masc.) and compounding (e.g. Pol. *leasingodawca* 'leasing donor'; Pol. *leasingobiorca* 'leasing recipient'; Pol. *auto-leasing* 'car leasing').

The tool used for the verification of morphological adaptation proves profitable in the assessment of the institutionalisation of various loan types, including for instance English abbreviations, which are inflected in Polish, e.g. Eng. *GPS* > Pol. *GPS-a/dżipiesa* [GEN.sg.], *GPS-owi/dżipiesowi* [DAT.sg.]; *GPS-em/dżipiesem* [INST.sg.]; *GPS-ie/dżipiesie* [LOC.sg.], and function as bases for adjectival, nominal and verbal derivatives (in graphic variants), cf. e.g. Pol. *piarowy/piarowski* (adj.) < Pol. *piar/PR* < Eng. *PR*; Pol. *ircownik* 'an IRC user' < Pol. *IRC* < Eng. *IRC*; Pol. *owatować/ovatować* 'to impose a VAT' < Pol. *VAT* < Eng. *VAT* (examples after Witalisz 2019). In most cases, predefining the potential graphic adaptation variant is necessary.

Loan translations from English may undergo morphological adaptation, yet the search for examples is far from automatic. The potential derived forms must be predefined and searched for as separate lexemes. Consider the following examples of univerbation, coined in Polish from multi-word loan translations from English: Pol. *sieciówka* < Pol. *sklep sieciowy* < Eng. *chain store*; Pol. *śniadaniówka* < Pol. *telewizja śniadaniowa* < Eng. *breakfast television*. Due to the resulting polysemy, corpus-assisted search for derivatives coined from semantic loans is a challenging task, though not impossible, cf. e.g. Pol. *jastrzębica* 'hawk-ess' [in a text about H. Clinton] and Pol. *jastrzębi* (adj.) 'hawk-like', both from Pol. *jastrząb* < Eng. *hawk* 'with a combative attitude'; Pol. *gołębi* 'dove-like' [in a text about B. Obama] < Pol. *gołąb* < Eng. *dove* 'advocating conciliation' (examples after Witalisz 2015).

3.5. Verification of lexical adaptation of multi-word loans and dialectal variants

The usefulness of corpora tools for the verification of the lexical adaptation of loans and their dialectal variants will be illustrated with multi-word loan translations from English, which, as evidenced by concordances found in the NKJP, serve as a starting point for lexical innovation. The adaptation of a multi-word loan translation may involve the substitution of one or more of its lexical components, as in Pol. *zamieść coś pod dywan* (< Eng. *to sweep sth under the carpet*), whose lexical elements frequently undergo substitution in Polish, e.g. the verb *zamieść* 'to sweep' happens to be replaced with *sprzątnąć* 'clear', *schować* 'hide', *ukryć* 'conceal', while the noun *dywan* 'carpet' is pragmatically replaced with *wycieraczka* 'doormat', *szafa* 'wardrobe/closet', *ołtarz* 'altar', etc. Pol. *Pierwsza dama* (< Eng. *First Lady*) and a series of other *First*-expressions loan translated into Polish resulted in lexical adaptations such as: Pol. *pierwszy obywatel* 'first citizen' [in reference to the Polish president], *pierwszy teść* 'first father-in-law' and *pierwsze dziecko* 'first child', the latter two used in reference to the relatives of a president.

Corpus tools make it possible to search for other cases of lexical adaptation of multi-word loans, such as modification, lexical extension, fusion, idiomatic allusion and disintegration, yet such a search is only semi-automatic and requires considerably more effort and time on the part of the researcher.

The dialectal variants of multi-word idioms loan translated from English may also be verified with corpus tools provided the researcher is familiar with the dialectal variants of the standard words used in loan translated expressions. Concordances have been found for two Standard Polish loan translations, *gorący ziemniak* (< Eng. *hot potato*) and *ziemniak kanapowy* (< Eng. *couch potato*), that have their dialectal variants with the noun *kartofel* (itself a German loanword, cf. Ger. *Kartoffel*) replacing *ziemniak* 'potato'.

3.6. Collocability attestation

The PELCRA search engine provides a tool called *Kolokator* for the retrieving of collocations and collocators. It is possible to select criteria such as part of speech, number of words, type of text as well as register and time span. A quick search for the query *leasing* tells us that it co-occurs in Polish with 485 other words and most often with the preposition *w* 'in' (1144), with the verbs *być* 'to be' (350) and *wziąć* 'to take' (292), and with the noun *samochód* 'car' (347), which corresponds to the well-established collocation *wziąć samochód w leasing* (lit. 'to take car in leasing']. The query results change significantly if in a more advanced search we check the collocability of the inflected and morphologically adapted forms of *leasing*, which still collocate most often with the

preposition *w* 'in' (2604), but also with the nouns *umowa* 'agreement' (1222), *firma* 'company'(1234) and the verb *być* 'to be' (1067).

The extraction of collocations is an important methodological tool in the study of covert loans, in particular in the case of semantic loans when no separate concordances are provided for the various meanings of a polysemous lexeme. The idiomatic use of Pol. *ciasteczka*, a semantic loan from Eng. *cookies* in its computer-related sense, can be attested through a careful semantic study of the collocators displayed by the collocation search engine for the word *ciasteczka*. It co-occurs most often in Polish with the preposition *na* 'for/on' (197), the adjective *kruche* 'crisp' (105), and the verbs *piec* 'to bake' (100) and *być* 'to be' (198). Its collocational potential with computer-related words is rather weak, cf. *przeglądarka* 'browser' (8), *Internet* (6), and *plik* 'file' (5), when compared to the lexical loan *cookie* that is also used in Polish and co-occurs with *plik* 'file' (198), *użytkownik* '[computer]user' (49), *ciasteczka* 'cookies' (17), and *przeglądarka* 'browser' (8). This tells us that of the two loan types coexisting in Polish, the loanword *cookies* is used more often than the semantic loan *ciasteczka*. This cannot be confirmed through the regular frequency query, since neither of the two search engines is capable of separating the non-idiomatic culinary sense of *ciasteczka* from its computer-related sense.

3.7. Verification of semantic development/reduction in loans

The concordances we obtain while searching for the frequency of loans and their morphologically adapted forms may be used for the studying of semantic changes that foreign loans undergo in the recipient language. Analysing the contexts of use provided by the corpus, we are able to compare the ways loans are used in the recipient language to their original meaning in the source language. Concordances found for Pol. *drink* (< Eng. *drink*) prove it is used only in one of the original senses of the English etymon, i.e. 'an alcoholic drink'. On the other hand, Pol. *sponsor* (< Eng. *sponsor*) is used to refer to 'a person offering financial assistance to another person in return for sex' alongside its original neutral English sense 'one that finances a project, event, or organization' (FD).

A long-term analysis of the concordances provided by the corpus lets us observe the semantic development of loans through time. Pol. *strefa zero*, loan translated from Eng. *Ground Zero* in September of 2001 and used in reference to WTC, was a case of semantic reborrowing in 2005 when it was used to refer to Bay St. Louis devastated by Hurricane Katrina, and a case of reinterpretation in the recipient language when used to label two local catastrophic events in 2006 (Witalisz 2015). More such extensions may be found in the corpus, e.g. Pol. *Happy hours*, borrowed directly from English, and Pol. *szczęśliwe godziny*, loan translated from Eng. *Happy hours*, have been found in Polish slogans advertising phone companies, Internet providers, beauty parlours, cheap

restaurants and second-hand shops, which contrasts with the use of the English etymon 'a period of time during which a bar or lounge offers drinks or food at reduced prices' (FD).

3.8. Verification of the co-existence of various loan types

One other asset of the corpus is the possibility to verify the co-existence of various loan types along with the time span over which this co-existence has occurred. It is not infrequent that multi-morpheme loanwords co-occur with their loan translated or loan blended versions over the same period of time. Pol. *fast food* (644 concordances plus 63 for *fastfood*) outnumbers its loan translated version *szybkie jedzenie* (63 concordances), just as Pol. *junk food* (44) exceeds in the number of concordances its loan translation variant *śmieciowe jedzenie* (26). This relation is maintained in several other cases, e.g. Pol. *Happy hours* (45 + 32 *Happy hour*) and *szczęśliwe godziny* (16), which shows the recipient language users' preference for unadapted loans.

In a more detailed corpus search we find evidence for the co-existence of a loanword *e-book*, a loanblend *e-książka* and a loan translation *książka elektroniczna*, all of which in their basic singular forms are attested in 346, 20 and 22 concordances, respectively. On occasion, the co-existing types of loans differ semantically, as in the case of Pol. *drapacz chmur* (222 concordances), an inexact loan translation from Eng. *skyscraper*, and Pol. *skyscraper* meaning 'an advertisement in the form of a web banner displayed vertically on a web page' (33).

The automatic verification of the co-existence of various loan types is not effective in the case of idiomatic loan translations that are homonymous to native loose syntactic phrases, as well as in the case of one- and multi-word semantic loans. In none of these cases is it possible to obtain separate concordances for the borrowed foreign senses, which will be elaborated on in the following section.

4. Limitations of corpus tools in the analysis of loans

While corpora tools, as evidenced above, come useful in verifying various formal aspects of the use of preidentified foreign loans in the recipient language, in more detailed studies, especially those of covert loans, corpus tools are not developed enough to provide separate semantic annotations for polysemous expressions or differentiate between coincidental homonymous pairs. In a corpus-based analysis of loans, language contact researchers have to resort to the manual extraction of contexts that meet the required semantic or formal criteria. In the following sections, we present some of the problems that appear in a corpus-assisted research on foreign loans.

4.1. Identification of anglicisms

Before we address corpus tools deficiency pertaining to semantic and structural analysis of loans, it must be stated that, as already claimed in the Introduction, corpora offer no tools for a fully automatic retrieval of anglicisms. Any corpus-based analysis of loans of English (or other) origin must be preceded with a manual loan identification by a language contact researcher, which requires theoretical background, research experience, and diachronic studies of dictionaries and language corpora of the two languages in contact. Manually preidentified loans may then serve as a starting point in a more detailed corpus-assisted search for and analysis of loan frequencies and other features described in Section 3.

Various attempts have been made at automatic identification of anglicisms, based chiefly on spelling recognition, through identifying the most typical grapheme n-grams in English words, i.e. 2- or 3- letter sequences characteristic of English and untypical of the recipient language, combined with machine learning methods and frequency-based strategies for the selection of features. Statistical methods were supplemented with a comparative lexicon-based method and regular word-formation rules (Furiassi and Hofland 2007; Furiassi 2008; Andersen 2005, 2011, 2012; Losnegaard and Lyse 2012). None of these methods used in isolation works effectively but if used complementarily, the level of precision for anglicism identification in Norwegian reaches between 60 per cent (Losnegaard and Lyse 2012: 150) and 75 per cent (Andersen 2011). Thus automatically retrieved anglicisms are still verified manually for identification correctness. The tools developed for Norwegian are unavailable for many other languages in which the influx of anglicisms is equally intensive. Therefore a search for anglicisms and their adaptation, usage and development in the recipient language, must be preceded with a ready list of anglicisms excerpted manually from traditional sources or semi-automatically from corpora if appropriate tools have been developed for a particular recipient language.

4.2. Search for lexical loans homonymous to native lexemes

Foreign loans happen to be graphically homonymous to native lexemes (or their infected forms). Such a co-existence may be a trap for a language contact researcher who wishes to investigate the frequency of a loan that is formally identical to the recipient language word. The NKJP search engines offer no tools that would be capable of separating the English loanword *baby* (< Eng. *baby*) from its Polish graphic homonym *baby* /babɨ/, which is the plural Nominative form of *baba* ‘*coll*. a woman’. The desired automatic search for the frequency of the English loanword has to be replaced with a manual extraction of concordances in which both words function as keywords. This seems

a daunting task in view of the 22,358 concordances that are displayed for *baby**
in the NKJP, and even for the 11,630 attestations found for *baby* without its in-
flected forms.

While it does not solve the frequency or morphological adaptation prob-
lem, to verify the very use of Eng. *baby* in Polish, it is useful to search for the
potential graphic adaptation of the loanword, which, following the Polish pho-
netic system, might be spelt *bejbi* /bɛjbi/. A quick search results in 91 concord-
ances, which seems an underrated value when compared to the attestations of
the English-sourced *baby*.

4.3. Verification of loan translations homonymous to native phrases

A similar problem concerns loan translations that happen to be homonymous
to loose syntactic phrases in the recipient language. No separate concordanc-
es that display the frequency or inflectional forms for each of these two types
can be obtained for the English-sourced idiomatic loan translations and native
word combinations that are formally identical to the former.

This is illustrated by numerous instances of idiomatic expressions loan
translated from English; for reasons of space we quote a few selected exam-
ples: Pol. *czarny kapelusz* < Eng. *black hat [hacker]*; Pol. *szklany sufit* < Eng.
glass ceiling 'discriminatory barrier', Pol. *czarny piątek* < Eng. *Black Friday*; Pol.
z tyłu głowy < Eng. *at/in the back of one's mind* 'present in one's thoughts';
Pol. *biały kołnierzyk* < Eng. *white collar [workers]*; Pol. *czarny koń* < Eng. *dark
horse* 'who achieves unexpected success', Pol. *gorący ziemniak* < Eng. *hot po-
tato* 'a controversial problem'. A similar problem occurs at the attempt to auto-
matically separate the idiomatic and literal uses of these English expressions
in a corpus of English.

4.4. Search for semantic loans in the corpus

Semantic loans are foreign meanings (senses) of words that have been added
to native vocabulary items. The old native meaning usually co-exists with the
new foreign sense, which results in the polysemy of the recipient language
lexeme. The NKJP search engines are incapable of separating concordances for
the different meanings of polysemous lexemes that have acquired new sens-
es under foreign influence (this cannot be done for native polysemes either).
Thus, automatic examination of the frequency or inflectional forms of English-
sourced semantic loans in Polish, such as e.g. Pol. *ciasteczka* < Eng. *cookies*,
Pol. *mysz* < Eng. *mouse*, Pol. *robak* < Eng. *bug*, Pol. *ikona* < Eng. *icon* in their
computer-related senses, is not possible. The same applies to other semantic
loans from English, e.g. Pol. *ekonomiczny* adj. < Eng. *economy* 'inexpensive', Pol.

lekki < Eng. *light* 'with less harmful ingredient', Pol. *jastrzębie i gołębie* < Eng. *hawks and doves* 'in business/politics-related senses', Pol. *aplikacja* < Eng. *application*, which has acquired several new senses from English: 1. 'computer programme', 2. 'job application', 3. 'the act of applying for', 4. 'applying a cosmetic onto the skin/hair', as well as to multi-word semantic loans, cf. Pol. *Koń trojański* in its computer-related sense (< Eng. *Trojan Horse*) and native *koń trojański* in its classical meaning.

The frequency of semantic loans can only be partially assessed with the use of the *Kolokator* tool that retrieves collocations and collocators, and indirectly confirms the various meanings of the analysed word.

4.5. Semantic disambiguation of polysemous loans

The deficiency of corpora tools in studying the polysemies of lexemes are also noticeable in the case of loanwords whose various meanings have been borrowed separately at different times by the recipient language users, as in the case of Pol. *grillować* (Eng. < *to grill*) that has been used in Polish since ca. 1998 to mean 'to broil on a gridiron' (the noun *grill* used since 1992), while around 2014 the English verb was reborrowed into Polish in the informal sense 'to question relentlessly; cross-examine'. In such cases, no separate concordances are displayed for the different meanings of polysemous anglicisms and semantic disambiguation must be carried out manually. This also means that no separate frequencies can be automatically obtained for each of the senses.

Semantic reborrowing is not infrequent in intensive and long-term language contact, and occurs also in the case of other types of loans, e.g. in loan translations and multi-word semantic loans, as in Pol. *przypudrować sobie nos* that extended its literal meaning under the influence of Eng. *to powder one's nose* 'to depart to the bathroom', and later was reborrowed to mean 'to use cocaine'. A researcher willing to examine the different uses and contexts of such cases of semantic reborrowing, as well as to verify the frequency and time of occurrence of a particular sense in the recipient language, must resort to manual browsing through the semantically unorganised concordances.

4.6. Verification of a foreign structural model

Finally, we move away from the semantic aspects of foreign loans and their low verifiability with corpus tools to examine the usefulness of the NKJP search engines for the measuring of the productivity of contact-induced word-formation patterns.

One of the manifestations of the intensive, mostly unidirectional, English-Polish language contact is the adoption of a Germanic word-formation rule for the derivation of compound nouns in Polish. N+N right-headed endocentric

compounding, typical of English, yet until recently unproductive in Polish, has become a productive word-formation rule, as claimed in Witalisz (2018). The fact that N+N compounding violates the grammatical laws of Polish does not impede its growing productivity, which is seen as a by-product of intensive lexical borrowing of N+N compounds from English, especially those in which one of the lexical elements reappears. Cf. English *business* in the following graphically adapted compound loanwords borrowed into Polish: Pol. *biznes class*, *biznes club*, *biznes lunch*, *biznesmen*, *biznes plan*, *bizneswomen*, which serve as models for Polish structural neologisms, such as for instance Pol. *biznes wiadomości* 'business news' (rule-governed Pol.: *wiadomości biznesowe*, N+ADJ), Pol. *Miłosz Festiwal* 'Miłosz Festival' (rule-governed Pol.: *Festwial Miłosza*, N+N$_{Gen.}$), Pol. *Wygoda-But* 'comfort shoe [brand name]' (rule-governed Pol.: *Wygodne Buty* ADJ+N).

To verify the productivity of the English-sourced word-formation rule, i.e. its application in composing native N+N right-headed endocentric compounds in Polish, we seek the help of a corpus tool, which, however, does not prove useable in this respect. The query: [pos~~subst & case=nom] [pos~~subst & case=nom], in which we define the part of speech as noun and limit the inflectionality of both nouns to the Nominative case, results in 1,454,040 (useless) concordances. Having analysed some of them lexically, we design a more elaborate query:[3]

[pos~~subst & cas=nom & !base="pan|pani|ksiądz|profesor|to|co" & !orth="[A-Z].+"]
[pos~~subst & cas=nom & !orth="[A-Z].+"]

in which we, additionally, limit the use of nouns and pronouns that appeared most frequently in constructions following the N+N structure, yet not having the status of compound words. Cf. for instance Pol. *Pan/Pani* 'Mr/Ms' that typically precedes another noun in phrases such as *Pan/Pani Minister* 'Mr/ Ms Minister'. Other lexical items excluded from the query include nouns: Pol. *ksiądz* 'priest' and Pol. *profesor* 'professor', as well as the pronouns Pol. *to* 'it' and Pol. *co* 'what'. Capital letters were also banned to eliminate phrases including proper names of the type *Pan/Pani X* 'Mr/Ms X'.

The nearly 30,000 concordances displayed for the second query were still not very helpful, as most of them included combinations that are typical of Polish, i.e. N+N$_{Gen.}$ of two types, in which the Genitive form of the second noun is identical formally (homonymous) with another noun in the Nominative case, as in a) Pol. *świst strzał* 'whizz [of] arrows': the isolated form *strzał* may be interpreted as either the Genitive case plural form of Pol. *strzała* 'arrow' or Pol. *strzał* (Nominative case, sing.) 'shot'; and b) Pol. *sprawa kobiety*

---

[3] With the assistance of Rafał Górski, Institute of the Polish Language, Polish Academy of Sciences, Kraków, Poland, spring 2018.

'case [of] woman': the form *kobiety* is either the Genitive case singular form of Pol. *kobieta* 'woman' or the Nominative case plural of the same noun, i.e. *kobiety* 'women'. The other concordances were cases of either misspellings of proper names (lower case) or combinations of two nouns in which the modifying noun followed its head, an unproductive word-formation pattern in Polish, as in Pol. *pszczoła robotnica*, lit. 'bee worker' (Eng. *worker bee*). The inflectionality of Polish and a high degree of inflection-caused homonymy make searches for N+N right-headed compounds counterproductive.

5. Conclusions

On the whole, despite the deficiencies described in Section 4, corpus search engines offer efficient and convenient tools for studying foreign loans, provided we search for expected features of preidentified lexical loans. This yields the first general observation that a corpus-assisted search for foreign loans is largely corpus-based, i.e. every search question is an expression of hypothesis about English loans, which can be verified and either confirmed and refined or disproved. The hypotheses are sourced in the awareness and knowledge of the grammatical complexities and phonological features of the recipient language, potential loanword formal adaptation processes, variety of loan types, and the existence of dialects in the recipient language.

Secondly, the factor that lowers the corpus tools efficiency is homonymy, which, as stems from the corpus study of foreign loans, is a frequent phenomenon, arsing, among others, from the high inflectionality of the recipient language. In the case of loan translations and semantic loans, composed of native material, corpus tools are (selectively) useful, yet the search is more time-consuming and calls for more expertise in corpus tools, especially in cases of polysemy and homonymy, which seem the most significant problems seeking a solution in corpus-assisted studies of foreign loans. Separate concordances can be obtained neither for English loanwords that happen to be formally identical to native Polish forms, nor for English-sourced idiomatic loan translations that coincide with homonymous Polish loose syntactic phrases, nor for semantic loans that are identical in form with native lexis. The dictionary lookup, a lexicon-based method for semi-automatic identification of lexical anglicisms described in Andersen (2011), might not be effective in cases of contact-induced homonymy.

Generally, corpus tools work more efficiently for loanwords than for covert loans composed of native lexical material. Loanwords which stem out from the recipient language texts due to a high degree of foreignness are most easily analysable with corpus tools that come useful in verifying various specific features such as frequency of use, formal adaptation, semantic development, derivative potential and collocability. While loanwords are relatively easily

detectable, even for non-specialists, semantic loans and loan translations occupy the other end on the recognisability scale due to their formal nativeness. In the middle of the continuum lie loanblends, whose automatic corpus search is effective, provided they are one-word or hyphenated expressions. Generally, as for the verification of loanword graphic adaptation and extraction of one-word loanblends and contact-induced hybrid compounds, semi-automatic extraction seems to bring satisfactory results.

Most problems with a corpus-assisted automatic search for loans described in this article are related to meaning, the least palpable and therefore the most difficult to describe and investigate aspect of language. A corpus-assisted search for and analysis of covert loans, semantic loans in particular, are much more time-consuming and much less effective than in the case of overt loans, which harmonises with the difficulties involved in the very identification of covert loans in the recipient language. To increase the usefulness of corpus tools for the study of semantic loans and loan translations, words would have to be annotated manually for their idiomatic senses.

Although at present language contact researchers have to rely on manual methods of extracting anglicisms, resent research shows that developing efficient data processing tools that will be successful in anglicims retrieval is only a matter of time (Renouf 2007; Andersen 2011, 2012). They will have to be language-specific, i.e. the grapheme typicality- and dictionary-based automatic extraction methods mentioned above will have to be language-specific and grounded on carefully selected morphological features and dictionaries compiled for particular receiving languages. Still, the automatic extraction tools are likely to be limited to non-adapted loanwords composed of overt source language material that displays foreign orthographic characteristics. Corpus-assisted lexicographic research in the field of contact linguistics necessitates further work on automatic recognition of word polysemy and homonymy.

References

Andersen Gisle (2005). Assessing algorithms for automatic extraction of anglicisms in Norwegian texts. In *Proceedings of the International Conference of Corpus Linguistics CL2005*. Birmingham: University of Birmingham. URL: https://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx [accessed June 27, 2019].

Andersen Gisle (2011). Corpora as lexicographical basis – the case of anglicisms in Norwegian. *VARIENG. Studies in Variation, Contacts and Change in English 6*. URL: http://www.helsinki.fi/varieng/series/volumes/06/andersen/ [accessed June 27, 2019].

Andersen Gisle (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In *The Anglicization of European Lexis*, Cristiano Furiassi,

Virginia Pulcini, Felix Rodríguez González (eds.), 111–130. Amsterdam/Philadelphia: John Benjamins.

Bańko Mirosław (ed.) (2000). *Inny słownik języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Dubisz Stanisław (ed.) (2003). *Słownik języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Furiassi Cristiano (2008). What dictionaries leave out: new non-adapted Anglicisms in Italian. In *Investigating English with corpora*, Aurelia Martelli, Virginia Pulcini (eds.), 153–169. Monza: Polimetrica.

Furiassi Cristiano, Hofland Knut (2007). The retrieval of false anglicisms in newspaper texts. In *Corpus linguistics 25 years on*, Roberta Facchinetti (ed.), 347–363. Amsterdam: Rodopi.

Görlach Manfred (ed.) (2001). *A Dictionary of European Anglicisms. A Usage Dictionary of Anglicisms in Sixteen European Languages*. Oxford: Oxford University Press.

Kurcz Ida, Lewicki Andrzej, Sambor Jadwiga, Woronczak Jerzy (1974–1977). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*. Vols. I–V. Warszawa: PAN, Instytut Języka Polskiego.

Kurcz Ida, Lewicki Andrzej, Sambor Jadwiga, Szafran Krzysztof, Woronczak Jerzy (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Vols. I–II. Kraków: PAN, Instytut Języka Polskiego.

Lewandowska-Tomaszczyk Barbara, Wilson Paul A. (2018). Sources of data and methodological foundations of a contrastive linguistic analysis of emotion concepts. *Bulletin de la Societe Polonaise de Linguistique* LXXIV, 157–189.

Losnegaard Gyri Smørdal, Lyse Gunn Inger (2012). A data-driven approach to anglicism identification in Norwegian. In *Exploring newspaper language. Using the web to create and investigate a large corpus of modern Norwegian*, Gisle Andersen (ed.), 131–154. Amsterdam/Philadelphia: John Benjamins.

Mańczak-Wohlfeld Elżbieta (2004) Does the spread of English constitute a threat to Polish? In *Speaking from the Margin. Global English from the European Perspective*, Anna Duszak, Urszula Okulska (eds.), 177–182. Frankfurt: Peter Lang.

Moon Rosamund (1998). *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Pędzik Piotr (2012). *Wyszukiwarka PELCRA dla danych NKJP. Narodowy Korpus Języka Polskiego*. Andrzej Przepiórkowski, Mirosław Bańko, Rafał Górski, Barbara Lewandowska-Tomaszczyk (eds.). Warszawa: Wydawnictwo Naukowe PWN.

Przepiórkowski Adam (2004). *The IPI PAN Corpus. Preliminary Version*. Warszawa: Institute of Computer Science, Polish Academy of Sciences.

Przepiórkowski Adam, Bańko Mirosław, Górski Rafał, Lewandowska-Tomaszczyk Barbara (eds.) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Przepiórkowski Adam, Górski Rafał, Łaziński Marek, Pęzik Piotr (2017). *Recent Developments in the National Corpus of Polish*. [pdf. pp. I–VII. ED July 18, 2017]

Renouf Antoinette (2007). Corpus development 25 years on: from super-corpus to cyber-corpus. In *Corpus linguistics 25 years on*, Roberta Facchinetti (ed.), 27–49. Amsterdam/New York: Rodopi.

Sᴄʜᴍɪᴅ Hans-Jörg (2010). Does frequency in text instantiate entrenchment in the cognitive system? In *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*, Dylan Gʟʏɴɴ, Kerstin Fɪsᴄʜᴇʀ (eds.), 101–133. Berlin/New York: Mouton de Gruyter.

Wɪᴛᴀʟɪsᴢ Alicja (2015). *English Loan Translations in Polish: Word-formation Patterns, Lexicalization, Idiomaticity and Institutionalization*. Frankfurt am Main: Peter Lang.

Wɪᴛᴀʟɪsᴢ Alicja (2018). Contact-induced right-headed interfixless N+N compounds in Polish. A corpus-based study. *Studies in Polish Linguistics 13*(1): 45–67.

Wɪᴛᴀʟɪsᴢ Alicja (2019). Polish faces of English acronyms and alphabetisms: An illustration of contact-induced linguistic diversity (Part 2), *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 136(1): 51–65.

Electronic sources:
FD – *The Free Dictionary*. https://www.thefreedictionary.com
http://korpus.ia.uni.lodz.pl
http://pelcra.pl
http://korpus.pwn.pl
www.ijp-pan.krakow.pl

Elżbieta Mańczak-Wohlfeld
Instytut Filologii Angielskiej
Wydział Filologiczny
Uniwersytet Jagielloński
al. Mickiewicza 9A, 31-120 Kraków
manczak(at)uj.edu.pl

Alicja Witalisz
Instytut Neofilologii
Wydział Filologiczny
Uniwersytet Pedagogiczny w Krakowie
ul. Karmelicka 41, 31-128 Kraków
alicja.witalisz(at)up.krakow.pl