Ewa Kozioł-Chrzanowska
PAN
Kraków
https://orcid.org/0000–0001–6308–7156

The Problems and (some) Solutions of Identifying Key Multi-word Expressions (MWEs). The Case Study of Polish Newspeak

Abstract. The paper aims to indicate and solve problems with practical usage of methods created for identifying key MWEs. The analysis is carried out on the basis of linguistic material representing Polish Newspeak (the language of propaganda and its mass media in totalitarian period). The paper considers three challenges: preparing an initial list of units which are supposed to be key ones, collecting searchable linguistic data and choosing the criteria of selecting appropriate texts. These problematic decisions which have to be made before analysis are inspired by works by Anna Wierzbicka and Raymond Williams (the notion of key MWEs is understood analogical to the key words in the interpretation of these authors).
Key words: *keyness, multi-word expressions, MWEs, Newspeak, keywords*

1. Introduction

1.1. The notion of key multi-word expressions (MEWs)

In linguistic studies, notions like *keyness* or *key words* (*keywords*) are understood in many different ways. According to Stubbs (2010), three loosely related, derived from different academic tradition uses of the term *keyword* can be indicated: words and culture (Williams 1976/1983), words and texts (Scott and Tribble 2006), phrases and schemas (Francis 1993) (Stubbs 2010: 23–32). The theoretical basis for the given paper is the first group, i.e. the interpretation derived from the cultural studies carried out by Williams or Wierzbicka[1];

[1] According to these authors, *keywords* are "(...) focal point around which entire cultural domains are organized" (Wierzbicka 1997: 156) "(...) significant, binding words in certain activities and their interpretation; they are significant, indicative words in certain forms of thought" (Williams 1975: 15).

the term *key multi-word expressions* is understood by analogy to the *keywords* in this sense. Generally speaking, *key multi-word expressions* are the ones which are focal, significant for the given culture or type of discourse.

1.2. Characteristics of the linguistic material

Presenting the analysis of identifying key MEWs of Polish Newspeak requires at least brief characteristic of the phenomenon. The term itself was coined by George Orwell in his *Nineteen Eighty-Four* to name an artificial, official language. The contemporary meaning is 'ambiguous euphemistic language used chiefly in political propaganda'. From 1944 to 1989 the Polish Republic was a non-sovereign country, dependent on the USRR as far as political and economic sense is concerned. It is claimed that during this period the official discourse was dominated by the Newspeak.

The Polish Newspeak has a few features, represented by groups of linguistic means. One of these features are pragmatic and semantic manipulations. They consist in giving new values to the language units, changing the components of their meaning, using the words with too general or too detailed meaning, e.g. *suggestions* was always used in a positive (e.g. *soviet suggestions*), while *declarations* – always in negative sense (e.g. *American declarations*), while there is no such division in the standard Polish. The next feature is using the schemata: conventional and repetitive phrases, metaphors and metonymies, e.g. *dalsze zacieśnianie braterskiej współpracy* (lit. continued bringing closer brotherly cooperation), *nierozerwalny sojusz* (lit. inseparable alliance). In the Newspeak the world is divided into two parts: "we" and "you". There is always an enemy, who is presented in an unambiguously negative way and has characteristic distinguishing marks like weakness and dispersion. Another feature is distortion. The aim of the Newspeak is to create the so-called information commotion. The information is usually incomplete, fragmentary or simply false. However, by using repetitive schemata, authorities create the impression of doing a lot of pivotal activities. This is the way of creating texts which are devoid of information, but are full of phrases showing the power of authority and its operations. Taboo is also characteristic for the Newspeak. Propaganda avoided some words and phrases by omitting or replacing them, e.g. instead of *strike* the *brakes in work* took place, political opponents were closed not in *prisons* but in *places of seclusions*. Another feature of Newspeak is simple syntax, stylistic monotony, lexical poverty (Markowski 2007: 90–94).

According to the Polish researchers, the Newspeak has a few functions: persuasive, distorting (disturbing), ritual, controlling functions as well as

functions consisting in making the so-called information commotion, manifesting the authority's presence and organizing social emotions (Markowski 2007: 87–90). Głowiński states that in Newspeak values dominate over the meaning. The language is subordinated to the rules of rituals, the magical thinking about it plays a pivotal role: the aim of the language is not to describe or to get to know the reality, but to create it by using words in a desired way (Głowiński 2001: 175).

1.3. Aims

The article aims to indicate challenges which come to light while using methods created for identifying key MWEs in practice. As a next step, the paper examines to what extent these methods can be useful with reference to the linguistic material of Polish Newspeak. The final interest of the article lies in providing some solutions to the indicated problems.

2. Analysis

2.1. First problem: an initial list

According to Scott (2009), methods of identifying keywords in texts can be divided into three main groups: relying on word frequency alone, basing on human identification and combinations of these both (Scott 2009: 2). It may be assumed that the same conclusions can be referred MWEs. All these groups are important scientific procedures and should be carefully judged in reference to the analyzed linguistic material. However, in the given paper, the methods based on human identification of key MWEs fall within the scope of the survey.

The well-known authors using these methods are probably Raymond Williams and Anna Wierzbicka. Williams identified the keywords intuitively, and then searched for empirical evidence of their historical shifts in meaning:

First, Williams identifies words intuitively, on the basis of his extensive scholarship. He then uses the attested citations in the 12-volume *Oxford English Dictionary* as empirical evidence that his keywords have undergone historical shifts in meaning which have led to complex layers of meaning in contemporary English. (Stubbs 2010: 23–24)

Wierzbicka claimed that keywords are very often centers of phraseological clusters and that they frequently occur in some special kinds of texts, like proverbs, sayings, songs and titles. The belief that these kinds of text

have special significance for the culture and that they reflect this culture is the basis for such an assumption:

> (...) one may want to show that this word is at the center of a whole phraseological cluster (...). One may also be able to show that the proposed "key word" occurs frequently in proverbs, in sayings, in popular songs, in book titles, and so on. (Wierzbicka 1997: 16)

It should be stated that the abovementioned methods based on human identification represent two different models. The first one – based on Williams' method – can be called an extraction model, because the action goes from data to the list of MWEs. In the second one (based on method by Wierzbicka), the direction of the action is opposite – that is why the method can be named as confirmation model. When Williams chose the keywords intuitively, he had some data – his own intuition, linguistic memory and competence which let him choose the words considered as key ones. This situation is analogical to automatic extraction of keywords from a corpus. The researcher has some data and as a result of the action he extracts from them a list of key MWEs. In the confirmation model the starting point is the list of MWEs. As a result of using a given method, the researcher gets the confirmation or rejection of the MWEs key status. Talking about methods of identifying key words or expressions is in fact a simplification, because some of these methods (representing the confirmation model) do not identify the words and expressions but confirm their key status. The distinction of two models is important, because using them incur slightly different practical problems.

The first challenge arising from the confirmation model is simply having the initial list of key MWEs. In other words, when the researcher wants to check if there are variants of expressions or if they occur in some kinds of texts, he needs to have these expressions first. Both, Williams and Wierzbicka, used their own intuition. Can this method be considered as a reliable one? To some extent, the answer may be positive. As Wierzbicka stated – if the researcher's choice is wrong, he will not get any interesting results. Some confirmation procedures can be used for checking intuitional choices. The bigger threat here is missing some important units. If the researcher omits them, they will probably remain omitted. Another problematic situation is having no or almost no intuition. In the case of Polish Newspeak, carrying out the research is problematic for those who do not remember the totalitarian period well or even at all. In such cases the only linguistic intuition about key words or MWEs of the past can be based on an idea built by books, films, newspapers and so on. The obvious advantage of older researchers is

their better intuition, which they could build by being immersed in the real, everyday various and live discourse.

In the case of Polish Newspeak, these problems are partially solved, thanks to works by Głowiński. Among his many books on the Polish Newspeak, there are four of them[2] in which the author describes particular words and phrases which he found crucial, interesting, surprising and so on. These books are a kind of linguistic diary – they were created on a regular basis in the totalitarian period of Polish history. They cover almost all years from 1966 to 1989. The tables of content of these books can be used as an initial list of key words and phrases of Polish Newspeak. Unfortunately, in the Polish scientific literature, we do not have any similar papers on the previous period (before 1966). The only possibility is to find some more general papers by different authors and note down the expressions which they describe. This piece of advide can be treated as a general solution to the problem of completing the initial list needed as a base for the research in the confirmation model. If the researcher looks for such a list, one of the possibilities is to search for the examples in as many various scientific works on the subject as possible. We can assume that their authors used plenty of sources or their own intuition, which is different from our own. The access to these works may be very helpful in the process of completing a list which can be processed in research being a part of a confirmation model.

2.2. Second problem: systematic search

The next indicated challenge of identifying key MWEs concerns both shown models (extraction and confirmation ones). No matter if the researcher wants to confirm his own assumptions that a given expression was a key one in a given period or if he wants to extract such phrases, he needs the collection of data which is searchable.

In the case of the extraction model, this data is available – first of all, the National Corpus of the Polish Language. One of the filters lets the users search only press texts, which seem to be the best source of propaganda (in comparison with books or the spoken language). The periods of publications can be also limited by choosing the years in which the newspapers were printed. The shortcoming of available search engines created for National Corpus is the lack of possibility of automatic extraction of a list of collocations. The same problem regards another source of texts – Chronopress,

[2] Głowiński 1991, 1993, 1996, 1999.

the portal of Polish press texts from year 1945 to 1954. However, this source is a part of the CLARIN – European Research Infrastructure for the Social Sciences and Humanities, focusing on language resources (data and tools). It means that we can easily use Chronopress with tools available in CLARIN, such as MeWeX, which is created for extracting the collocations from corpora.

The situation is much more complicated in relation to the confirmation model. In the abovementioned quotation Anna Wierzbicka (Wierzbicka 1997: 16) mentions proverbs, sayings, popular songs and book titles as the texts which are important for confirming the status of keywords. This list can be easily extended, e.g. to posters or internet memes. However, the pivotal problem lies in the accessibility of the data and the possibility of searching them. There is no collection of such texts which would let the researcher easily look for a word or expression in popular songs, for example. It is impossible to search this kind of texts in the same way in which the corpora can be analyzed. There are two main possibilities of solving this problem first – search many scattered sources, second – assume that many of these texts are available on the Internet and use its standard search engines to do the research.

2.3. Third problem: selection of texts

The last of the indicated problems is the challenge of selecting appropriate texts as the basis of the research. The quoted method by Anna Wierzbicka (Wierzbicka 1997: 16) assumes that some texts – which can be called "significant texts" – have special importance.

This importance is based on two mechanisms. Either the MWEs is key, so it appears in significant texts or the text is so significant that it makes MWE a key one. These two types of relationship are represented by illustrations 1 and 2.

The first illustration is the example of a significant text. The propaganda poster from 1945 became a symbol of a communist terror and postwar persecution of soldiers from Home Army. These soldiers are compared to dwarfs, as the communist propaganda was accusing them of collaboration with Germans and objection to social reforms. This metaphorical comparison in completely absent in other propaganda texts – there in no sample of this expression in Chronopress corpus (the abovementioned collection of press texts printed from 1945 to 1954). However, the poster is so significant and well-known that the MWE created for it became key. The opposite situation is visible in illustration no. 2. The internet meme uses the slogan *Polak potrafi* (lit. A Pole can). This expression was used on the building site

Illustration 1. The propaganda poster *Olbrzym i zapluty karzeł reakcji* (lit. The giant and the spat dwarf of the forces of reaction)

Source: the Internet.

Illustration 2. The Internet meme *Polak potrafi* (lit. A Pole can), example 1

Source: the Internet.

of Ironworks Katowice – a huge venture which became a symbol of Edward Gierek's (the first secretary of the communist party) era. Its meaning can be described as 'a Pole is a smart person who can deal with many problems'. The slogan became very popular and it is still used nowadays, usually in ironic contexts (like in the abovementioned meme, where it is a caption of the absurd construction being a mix of a car and a tractor). It is a key MWE of the Polish language and that is why it is frequently used in the internet memes.

At the same time the given illustrations represent two other categories of texts: the ones that are special by their exposition (a poster, no. 1) and the others which represent counterspeech[3] (an Internet meme, no. 2). Both categories are useful as data for extracting key MWEs or confirming their key status.

Some types of texts are constructed in a way that exposes some content. To this idea refers, among others, the notion of text clusters (Püschel 1997) which is used for example in the keywords analysis on the Internet. Commercials and press where titles, leads and covers of newspapers play a special role work according to similar rules. In order to check the importance of exposition factor in propaganda press, the analysis based on the Chronopress corpus was conducted. The research compared frequency of using the words in the newspapers in general and on their covers in 1945. Taking into consideration 1,000 most popular examples has shown that approximately 16 per cent of words most popular in general were not comparably popular on the covers of newspapers. In about 5 per cent the difference in popularity was bigger than 1,000 positions on the frequency list. For example *a church* was 586th most frequently used word in Polish press in 1945, but at the same time it was only 2489th on the covers. Such examples suggest that the frequency of word is not crucial. Can it be stated that the word or phrase is key if it is not exposed? Probably the answer should be negative. In other words, the fact that some words or expressions appear in the texts or parts of texts which are well exposed proves their key status (in the case of press the best exposed part of the text is definitely the cover).

The next criterion helpful in choosing the types of significant texts is their affiliation to counterspeech. It seems to be obvious that all examples of counterspeech are based on units which are well-known, established in a language, used at least by a small group of people. Otherwise, making

[3] Counterspeech is a linguistic phenomenon of opposing the traditional forms of communication used in a given society at a specific time, e.g. antiproverb.

Illustration 3. The Internet meme *Polak potrafi*, example 2 (lit. A Pole can)

Source: the Internet.

Illustration 4. The Internet meme *Polak potrafi* (lit. A Pole can), example 3

Source: the Internet.

counter-units based on them would not have any sense at all. Moreover, a large number of variants is a sign of an important role the unit plays in a language and culture. It can be stated that a MWE being a base for many various counter-units is key itself. Illustrations from 3 to 5 are the internet memes based on the slogan *a Pole can*. They constitute only a small sample of the collection which can be easily found on the Internet. Their number and variability are a sign of the fact that this slogan is important for Polish Newspeak, at least from the contemporary perspective.

Illustration 5. The Internet meme *Polak potrafi* (lit. A Pole can), example 4

Source: the Internet.

3. Conclusion

The aim of the paper was to indicate challenges of identifying key MWEs of Polish Newspeak and provide at least some solution to these problems. The first problem – creating an initial list of key MWEs of Polish Newspeak – has already been partially solved by Michał Głowiński's works. The period which was not described by the author needs a list created separately. This goal can be achieved by searching examples from different scientific works devoted to the topic of totalitarian propaganda. The second problem – systematic search in significant texts like sayings, songs and so on – can be solved either by using the Internet, or by a detailed enquiry of various sources. The third of the indicated problems is the challenge of selecting appropriate texts as the basis of the research. The solution can be provided on the basis of two main criteria: exposition and counterspeech. According to them, the texts which are well-exposed and/or represent counterspeech are significant enough to establish a collection of texts useful for searching of key MWEs.

Literature

Głowiński, Michał. 1991. *Marcowe gadanie 1966–1971*. Warszawa: PoMost.
Głowiński, Michał. 1993. *Peereliada 1976–1981*. Warszawa: PIW.

Głowiński, Michał. 1996. *Mowa w stanie oblężenia 1982–1985*. Warszawa: OPEN.

Głowiński, Michał. 1999. *Końcówka 1985–1989*. Kraków: Wydawnictwo Literackie.

Głowiński, Michał. 2001. Nowomowa. In: Bartmiński, Jerzy (eds.). *Współczesny język polski*. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej. 173–183.

Markowski, Andrzej. 2007. *Kultura języka polskiego*. Warszawa: PWN.

Püschel, Ulrich. 1997. Puzzle-Texte – Bemerkungen zum Textbegriff. In: Antos, Gerd; Tietz, Heike (eds.). *Die Zukunft der Textlinguistik. Traditionen, Transformationen*, Trends, Tübingen. 27–41.

Scott, Mike. 2009. Problems in investigating keyness, or clearing the undergrowht and marking our trails. In: Bondi, Marina; Scott, Mike (eds.): *Keyness in Text*. Amsterdam: John Benjamins. 43–59.

Stubbs, Michael. 2010. Problems in investigating keyness, or clearing the undergrowht and marking our trails. In: Bondi, Marina; Scott, Mike (eds.). *Keyness in Text*. Amsterdam: John Benjamins. 21–43.

Wierzbicka, Anna. 1997. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, Japanese*. New York Oxford: Oxford University Press

Williams, Raymond. 1975. *A Vocabulary of Culture and Society*. Oxford: Oxford University Press.

Internet sources

ChronoPress: Chronologiczny Korpus Polskich Tekstów Prasowych [Chronological Corpus of the Polish Press Texts]. http://chronopress-test.clarin-pl.eu/.

MeWeX. https://mewex.clarin-pl.eu/.

NKJP: *Narodowy Korpus Języka Polskiego* [National Corpus of the Polish Language]. http://nkjp.pl/.

Examples 1–4 – found by the Google Images search engine (as a result of searching every single MWE).

Los probemas y (algunas) soluciones para identificar expresiones multipalabras clave. El case de estudio: la Neolengua Polaca

Resumen

Indudablemente, algunas expresiones creadas por la Neolengua Polaca (el lenguaje creado por la propaganda y los medios de comunicación masiva en el periodo totalitario) están aún en uso en el idioma polaco (esto es después de la caída del comunismo y la República Popular Polaca). Su presencia en el lenguaje actual, así como sus cambios semánticos y pragmáticos, contribuyen a importantes problemas en el idioma y la cultura polaca. Sin embargo, un análisis detallado de estos problemas requiere responder a una pregunta básica: ¿Cuales de estas expresiones, a las

que llamamos MWEs, pueden ser reconocidas como palabras clave? Diferentes criterios para identificar estas palabras clave están seńalados en diferentes textos y son analizados basándose en publicaciones de la época totalitaria en Polonia. Como resultado, este trabajo muestra los problemas fundamentales y ofrece algunos soluciones a ellos. Las conclusiones pueden ser útiles para el caso de estudio que aquí se considera (identificación de palabras claves MWEs en el Neologismo Polaco), así como para otros textos del mismo perfil.