The Status of Equivalents in Bilingual Dictionaries and in a Language

Iwona Szlanszok
41-909 Bytom, Warzywna 16
Poland

Abstract
The paper endeavors to set out the essence of equivalence in bilingual English-Polish dictionaries. The basis for our research will be the Polish version of *Cambridge International Dictionary of English* (CIDEP). In the course of the research, Polish equivalents in two medium-sized English-Polish dictionaries will be compared with CIDEP. *Nowy słownik angielsko-polski* (T. Piotrowski, Z. Saloni) published in 2002 by Wilga and *Słownik podręczny angielsko-polski* (J.J. Kałuża) published in 2000 by Exlibris have been chosen for our study. In our research, we will focus on the Polish language of equivalents. We will examine whether the equivalents in dictionaries are close to the contemporary corpus of the Polish language. In our research, we have looked at the formal aspect of equivalents (the length of words and their frequency) as well as the meaning aspect on the basis of the word *iść*.

1 Introduction

In this paper we will analyze the Polish version of *Cambridge International Dictionary of English-Polish* (CIDEP) edited by Anna Duszak et al. (2003). In the course of the research, Polish equivalents in *Nowy słownik angielsko-polski* (Piotrowski, Saloni, 2002) and *Słownik podręczny angielsko-polski* (Kałuża, 2000) will be compared with those in CIDEP.

CIDEP contains almost 40 000 headwords, over 60 000 entries and over 87 000 lexical units. It is a monodirectional dictionary addressed to Polish speakers with an intermediate to advanced level of English. According to the editor, each English headword has a large number of Polish equivalents as large as possible. Polish equivalents range from the most important and widely used to those less common and rare. We will try to examine whether Polish equivalents used in the bilingual dictionaries are compatible with the contemporary Polish language. We will also try to answer the question: What is the status of the Polish equivalents in the Polish language?

2 Definitions and tools

The tool used for the automatic analysis is the program *Oxford WordSmith Tools 4.0* (M. Scott, 2005). *Oxford WordSmith Tools* is an integrated suite of programs: the Wordlist tool makes it possible to see a list of all the words or word-clusters in a text, the Concord tool enables one to see any word or phrase in context, and with the KeyWords tool the key words in a text can be found.

A **wordlist** is a list of all the word types. A **type** is a collection of tokens, and a **token** is an individual word occurring in a text. In our analysis, we will also deal with **type/token ratio** (TTR), which is simply the ratio of word types to word tokens in a corpus. The **standardized type/token ratio** (STTR) is computed every 1,000 words (Scott, 2005).

In order to compare the equivalents with the Polish language, we have chosen a collection of contemporary Polish texts from newspapers, periodicals, and magazines. We call this collection a **normalized text.**

3 The structure of CIDEP

For our study purposes the electronic version of CIDEP is used. It is encoded in XML (eXtensible Markup Language), a document interchange format. A dictionary entry of CIDEP in XML:

<haslo nazwa="**babel**"><sylaby>**ba**<bm />**bel**</sylaby><transkrypcja>**"beI;bXl**</transkrypcja>
<znaczenie><opis_gram>**rz zw. l poj**</opis_gram><kwalifikator>**przen**</kwalifikator>
<ekwiwalent>**wieża Babel**</ekwiwalent><ekwiwalent>**zgiełk**</ekwiwalent>
<ekwiwalent>**chaos**</ekwiwalent><ekwiwalent>**pomieszanie języków)**</ekwiwalent>
<grupa_przykladow><przyklad><tresc>**Communication between different computers has been made difficult by the babel****of****computer languages used by different machines.**</tresc>
</przyklad></grupa_przykladow></znaczenie>

Nowy Słownik is written in TEX (typesetting program) and tagged; and *Słownik Podręczny* in RTF (rich text format). For our purposes, the headwords along with their equivalents will be isolated from the dictionaries. Moreover, we will convert the database into the Polish-English index so as to clearly see the Polish equivalents.

4 Polish equivalents in bilingual dictionaries

In order to make a comparison with the contemporary Polish language we have included in our study a normalized text. The first observation is that the type/token ratio is quite high in the dictionaries in comparison with the normalized text. The same strings of words occur very rarely in the normalized text, which means that dictionaries have a rich vocabulary in their language of equivalents.

tokens (running words) in text		360 095
types (distinct words)		58 112
type/token ratio (TTR)		16
standardized TTR		56
mean word length (in characters)		5
	number of occurrences	% of all tokens
1-letter words	30 695	8.52
2-letter words	38 842	10.79
3-letter words	41 646	11.57
4-letter words	34 643	9.62
5-letter words	41 689	11.58
6-letter words	40 634	11.28

6-letter words	40 634	11.28
7-letter words	34 455	9.57
8-letter words	28 487	7.91
9-letter words	23 571	6.27
10-letter words	15 685	4.36

Figure 1. Statistical data of the Polish normalized text.

Figure 1 shows that in normalized Polish texts there is a characteristic decrease in the number of 4-letter words and the number of 5-letter words is the largest. Generally, there are more strings of words that have fewer than 6 letters.

tokens (running words) in text		152 352
types (distinct words)		44 825
type/token ratio (TTR)		29.45
standardized TTR		42.04
mean word length (in characters)		7.42
word length standardized		5.06
	number of occurrences	%of all tokens
1-letter words	4 427	2.91
2-letter words	4 141	2.72
3-letter words	8 931	5.86
4-letter words	7 268	4.77
5-letter words	15 335	10.07
6-letter words	18 517	12.15
7-letter words	19 353	12.70
8-letter words	19 269	12.65
9-letter words	17 589	11.54
10-letter words	14 126	9.27

Figure 2. Statistical data of Polish equivalents in CIDEP.

The data in Figure 2 show the decrease in 4-letter words, but on the other hand there are many words of more than 6 letters CIDEP. Now, let us look at the data of the other, smaller dictionaries.

tokens (running words) in text		52 607
types (distinct words)		20 773
type/token ratio (TTR)		39,99
standardised TTR		74,26
mean word length (in characters)		6,65
word length standardized		3,05
	Number of occurrences	% of all tokens
1-letter words	1 789	3.44
2-letter words	2 625	5.05
3-letter words	4 948	9.51
4-letter words	3 063	5.89
5-letter words	7 082	13.62
6-letter words	6 503	12.51

7-letter words	34 455	9,57
8-letter words	28 487	7,91
9-letter words	22 571	6,27
10-letter words	15 685	4,36

Figure 3. Statistical data of Polish equivalents in *Nowy Słownik.*

Figure 3 shows a medium-sized dictionary which also displays this characteristic decrease in the number of 4-letter words. Just as in CIDEP, we can observe many long words with the highest number of 5-letter words.

tokens (running words) in text		30 930
types (distinct words)		17 132
type/token ratio (TTR)		55,40
standardised TTR		60,99
mean word length (in characters)		7,40
word length standardised		2,74
	Number of occurrences	% of all tokens
1-letter words	339	1,10
2-letter words	504	1,63
3-letter words	1 602	5,18
4-letter words	1 691	5,47
5-letter words	3 507	11,34
6-letter words	4 314	13,95
7-letter words	4 444	14,37
8-letter words	4 265	13,79
9-letter words	3 577	11,56
10-letter words	2 689	8,69

Figure 4. Statistical data of Polish equivalents in *Słownik Podręczny.*

The smallest dictionary used in the study does not have this decrease of 4-letter words (see Figure 4) but, just like in the other dictionaries we can observe a large number of long words not so often used in normalized text. Moreover, when we look at the mean word length we can see a striking difference from 5-letter words in normalized text to an equal number of 7-letter words in bilingual dictionaries.

Now, let us examine another aspect of equivalents apart from their length. We can observe certain parallels with Polish normalized text in terms of word frequency. As Figure 5 shows the most frequent words in Polish are prepositions (*w, na, z, do*), pronouns (*się, to*), conjunctions (*nie, i, że*) and the verb *be* in one of its forms (*jest*).

N	Word	Freq.	%
1	W	9 580	2,66
2	I	7 583	2,11
3	SIĘ	7 229	2,01
4	NIE	6 731	1,87
5	NA	6 282	1,74
6	Z	5 732	1,59

7	TO	4 396	1.22
8	DO	4 146	1.15
9	ZE	3 394	0.94
10	JEST	3 003	0.83

Figure 5. The most frequent words in Polish normalized text.

Equivalents in English-Polish dictionaries are very similar in their frequency as it can be seen in Figure 6.

N	Word	Freq.	%
1	SIĘ	7 608	0.58
2	W	3 472	0.27
3	DO	2 886	0.22
4	NA	2 827	0.22
5	Z	2 583	0.20
6	NIE	1 052	0.08
7	COŚ	881	0.07
8	O	832	0.06
9	I	728	0.06
10	CZEGOŚ	654	0.05

Figure 6. The most frequent equivalents in bilingual dictionaries.

We also examined the most common Polish nouns and verbs in bilingual dictionaries. Among the most common nouns are *osoba* (person) freq. 448, *miejsce* (place) freq. 219, *człowiek* (human being) freq. 175, *czas* (time) freq. 130, and *praca* (work) freq. 114. As far the verbs are concerned, the research has shown that the most frequent ones are *być* (be) freq. 572, *mieć* (have) freq. 433, *robić* (do) freq. 320, *dawać* (give) freq. 174, *iść* (iść) freq. 147.

Now let us examine the verb *iść* with all of its concordance instances. To further illuminate the data, we have put the examples in the table (Figure 7) along with their English translation as it appears in the three English-Polish dictionaries.

	POLISH EQUIVALENT	*CIDEP*	*NOWY SŁOWNIK*	*PODRĘCZNY*
1	iść	go, walk, march, head, pass	walk	go, walk
2	iść dalej	follow on,push on	carry on, continue, get on, procede	go on
3	iść spać	turn in	retire, go to bed	turn in
4	iść w góry	go up, lift off, ascend	climb, move up	soar
5	iść za	go after	follow	Follow
6	iść czyimiś śladami	X	follow in sb's footsteps	X
7	iść fałszywym tropem	X	be barking up the wrong tree	X
8	iść po omacku	X	muddle through, grope	X
9	iść ociężale	X	X	slouch
10	iść zygzakiem	X	X	zigzag

Figure 7. Lexeme *iść* in the definition language of the English-Polish dictionaries.

In *Nowy Słownik* quite a large number of Polish collocations and phrases with *iść* are covered (44 instances) even though it is only a medium-sized dictionary. In CIDEP 56 phrases with *iść* are presented and in comparison with *Nowy Słownik* this is not such an impressive number for a large dictionary. Apart from this, *Nowy Słownik* gives quite sophisticated phrases that are not mentioned in CIDEP, like *iść (o sztuce), iść czyimiś śladami, iść fałszywym tropem, iść jak burza, iść jak cholera, iść jak woda, iść na całego, iść na opak, iść na skróty, iść po omacku, iść wbrew naturze, iść z płomieniami.* These phrases refer mostly to figurative meanings of *iść*.

Słownik Podręczny is the smallest dictionary, and in proportion to its size it covers the smallest number of phrases. But still, we can find here some equivalents not mentioned in any other of the two other dictionaries, such as *iść mozolnie naprzód, iść ociężale, iść powolnym krokiem, iść zygzakiem.* It seems that in *Słownik Podręczny* a manner of walking is emphasized in the equivalents.

5 Concluding remarks

In many cases, the equivalents appear to be chosen out of lexicographers' heads and based on their everyday knowledge of the language itself. In terms of the length of equivalents, we can even risk the statement that the essence of equivalence in bilingual dictionaries lies in sophisticated distinction between equivalents, since there are so many long words, unlike in the normalized text. Hence, such dictionaries become an awkward way of putting simple messages across to users. The key solution seems to be a more rational way of handling foreign words in the mother-tongue environment. To put it differently, when making a dictionary one has to focus on equivalents that have high scores in a target language corpus instead of choosing highly sophisticated types of equivalents.

References

A. Dictionaries

Duszak, A. (ed.) (2003), *Cambridge English-Polish Dictionary.* (First edition.), Warszawa, Prószynski i S-ka. (CIDEP)

Kałuża, J.J. (2000), *Słownik podręczny angielsko-polski.* (First edition.), Warszawa, ExLibris.

Piotrowski, T., Saloni, Z. (2002), *Nowy słownik angielsko-polski.* (Third edition.), Warszawa, Wilga.

B. Other literature

Atkins, B. T. S. (1993), 'Theoretical lexicography and its relation to dictionary making', in *Dictionaries: the Journal of the Dictionary Society of North America,* pp. 4-43.

Piotrowski, T. (1989), 'Monolingual and bilingual dictionaries: Fundamental differences', in Tickoo, (ed.) 1989, pp. 72-83

Piotrowski, T. (1994), *Problems in bilingual lexicography.* Wrocław, Wydawnictwo UW.

Scott, M. (2005), *WordSmith Tools,* Oxford, OUP.