Parallel Corpora as a Source of Defining Language-specific Lexical Items

Dmitri Sitchinava
Institute of Russian language, Russian academy of Sciences
e-mail: mitrius@gmail.com

Abstract

The paper presents an attempt to propose an exact method for identifying the so-called "language-specific" lexicon, a controversial notion often reasonably questioned. An aligned bilingual parallel corpus is chosen as an instrument for finding "specificity", and statistical entropy and other indices are used as markers of the dispersion of translation patterns (viz. stimuli). For example, a word can be deemed (maximally) language-specific if it occurs multiple times in a given bilingual corpus and is translated each time in a different way. A word is minimally (or simply not) language-specific if it is translated each time identically. Some problems relative to the application of this method are discussed. These data can be explicitly used in bilingual dictionaries.

Keywords: parallel corpora; language-specific lexicon; translation patterns; statistics

1 The Issue of Language-Specific Lexical Items

The question discussed in this paper is the one of alleged "language-specific" lexicon. These are items with no well-defined semantic counterparts in (at least some or even the majority of) other languages: such elements are often borrowed into other languages rather than translated, like German *Angst*, French *ennui* or Russian *intelligentsia*. Alongside with this, a group of authors tend to see in these facts, in a Neo-Humboldtian vein, a particular "linguistic view of the world" with such a lexicon as its core element; these items are proclaimed "key words" or "key concepts" of this alleged view (cf. for example, Зализняк et al. 2012). The term itself, German *Weltbild* (Rus. *kartina mira*), has been long known in different philosophical traditions, but applied to linguistics by a Neo-Humboldtian, Leo Weisgerber. Sometimes the "linguistic view of the world" is directly identified with the lexical-semantic system of a given language.

These ideas are widespread in the post-Soviet countries (especially Russia, where hundreds of dissertations and thousands of papers on "concepts", "language-cultural studies" and "linguistic view" are routinely written, but also, for example, Ukraine), Central/Eastern Europe, Japan, and within the Wierzbicka school in Australia. They are also not unknown in the West where they stem from the heritage of Boas, Sapir and their anthropological school, mainly Whorf's hypothesis of linguistic relativity. The Western reception of Whorfianism (Lucy 1992; Levinson 2003; Boroditsky 2011 and others) is far from being straightforward and is much more cautious; the "cultural" mechanism is by no means viewed by these authors as a universal one, and they search links rather between grammar and behaviour than between lexicon and culture.

The neo-Humboldtians and their followers are often severely criticized for an alleged nationalistic or anachronically «romantic» stance (Sériot 2005; Павлова (ed.) 2013; cf. an overview of the controversy, Рycco 2014). Indeed, the vulgarizations of these ideas (mainly in the post-Soviet space) are often used for overtly nationalistic and chauvinistic speculations, both by epigone linguists and politicians (e. g. «the words meaning ‘conscience‘ and ‘justice‘ are not well translated from the language X, therefore the people speaking X is more “spiritual” and has a superior mentality»). Even the best representatives of this school sometimes pronounce judgments on “national mentality” based on rather subjective philosophical essays that are not scientific findings.

The main purely linguistic (non-ideological) controversy is centred on the fact that the alleged language-specific lexical items do have, in fact, their counterparts in a given context, and that every sense behind them is indeed well translatable (see, in particular, Павлова (ed.) 2013). The “linguistic view” or “relativity” hypotheses are also criticized because they do not separate language from cognition and themselves rely on other, subjective hypotheses and definitions of cognitive processes; alongside with this, all independent judgments on languages become impossible by definition.

The author adheres to the point of view that there seems to exist no comprehensive “linguistic view of the world” interlinking all the unique points of the lexical system of a given language. I am also sceptical about the hypothesis that there exists an “ethnic/national mentality”, let alone reflected by the language of a given ethnos/nation. However the cross-linguistic study of lexical meanings is a legal and unquestioned branch of semantics and typology, and this trend of linguistics can well be interested in a cross-linguistic definition of “language-specific” lexical items without discussing the puzzle of “linguistic view” or “relativity”. We may quote here the famous lines by Vladimir Nabokov (“Nikolai Gogol”, 1944) on Russian *poshlost’* (‘~platitude, vulgarity’): “The absence of a particular expression in the vocabulary of a nation does not necessarily coincide with the absence of the corresponding notion but it certainly impairs the fullness and readiness of the latter’s perception. Various aspects of the idea which Russians concisely express by the term *poshlost* <…> are split among several English words and thus do not form a definite whole”. We are interested exactly in the phenomenon of “splitting” of the same word among different words of another language, and not in the “absence of the corresponding notion”.

It seems that an attempt at defining language-specific items can be made with less impressionistic and more exact means, viz. using statistics based on parallel corpora that include translated texts. The items in questions are not thus (completely) untranslatable; indeed, they ought to be and therefore are already translated (or, vice versa, themselves chosen via translation), and our task is to view the range of possibilities that emerges there. This result can also have consequences on compiling bilingual and even monolingual explanatory dictionaries.

2 Parallel Corpora as a Tool

Parallel corpora are pre-existing (not elicited or translated for the purpose of linguists) original and translated texts, typically aligned sentence-by-sentence. An original text can be aligned against more than one translation into different (and/or even the same) language. There exists a large amount of literature on parallel corpora. I may notice there Добровольский 2009 on “cultural lexicon” in a Russian-German parallel corpora (particularly different translations of Dostoevsky) and especially the book on the German lexicon Добровольский 2013 with a large section on parallel corpora. A special issue of the journal STUF (Cysouw, Wälchli 2007) is dedicated to multilingual corpora also known as “massive parallel texts” where we have one and the same text translated into dozens and even hundreds of languages. These corpora are a very impressive typological tool but their reach is naturally restricted by the texts that are read and propagated in many different countries (the Bible being the best translated one).

The Russian National Corpus (RNC, http://ruscorpora.ru) comprises different bilingual parallel corpora (http://ruscorpora.ru/search-para.html) where either the source language or the target language is Russian. Their amount is representative enough for studying lexicon and such elements as alleged language-specific items. For example, the bilingual English-Russian corpus includes 22 million words, the Ukrainian-Russian one, 9 million, and the German-Russian, 7 million. These corpora mainly include fiction, which is good for our purposes, as, for example, lexemes describing subjective feelings are less likely to be found in journalism or legalese. The translations are made by professional translators of fiction within the narrower and broader context: they are more informative than using a traditional dictionary where contexts and counterparts can be artificial.

Naturally there exist some methodological issues that are to be addressed in the study. For example, the translators are not necessarily fully bilingual: they can make overt mistakes and they can misconceive the meaning of a foreign lexeme as exactly the same as that of some native one (whereas some subtleties may be present). They can (and they indeed do) use bilingual dictionaries: while doing so they may copy, mechanically or not, the translation counterparts from these dictionaries, and this strategy can compromise these translations as an independent source. A single translator can choose once and forever a given counterpart for a given word and then insert it into all the contexts in all the books s/he translates (the case, say, of Constance Garnett, the great pioneer of translating Russian classics into English who was not fluent in Russian). To avoid some of these problems, the corpus should be large and representative as far as different authors and translators are concerned.

3 Dispersion/Entropy Indices For Lexemes

It is expected, as a zero hypothesis, that a common element of a lexicon that has a semantically neutral counterpart, like ‘cat’ or ‘table’, would have one or at least few well-defined counterparts. If we analyze a distribution of these counterparts, they are expected to have lower entropy of distribution:

$$H\ (M) = (\ —\ \Sigma(F(M_i)\ /F(O))\ \log_2(F(M_i)\ /F(O)))$$

and a bigger «monopoly index», such as Herfindahl number used originally in economics:

$$Herf\ (M) = \sum(F(M_i)\ /F(O))^2$$

There exists also a normalized Herfindahl index that ranges between zero and 1:

$$HerfNorm\ (M) = (Herf\ (M)\ —\ 1/NumM)\ /(1\ —\ 1/NumM).$$

F(M) and F(O) stand for the frequency of a given translation model and for the frequency of the word in question within the original text; NumM stands for the number of different translation models (patterns) for this word. The same indices can be calculated for the translated texts and different translation stimuli in the original texts as well; in the corresponding formulae, M(odel) and O(riginal) are substituted by S(timulus) and T(ranslation).

Conversely, to qualify as a language-specific, a word should be translated in a more dispersed way and have greater entropy, e.g. when it is attested *n* times in Russian it can be translated into English differently each time. It is also expected that a language-specific word can be more freely used in an original text than in a translation. We see that in this understanding language-specificity is a scale or even a set of scalar variables, rather than a yes/no value.

Note that we expect that a language-specific word would normally occur more frequently (in terms of instance per million) in an original text than in a translated one.

Some words usually claimed to be Russian language-specific (*prostor* ‘~space‘, *toska* ‘~yearning‘, *udal’* ~bravado‘, *ujut* ‘~cosiness‘, *poshlost’* ‘~platitude‘; cf. Зализняк et al. 2012, or the classics of this linguistic trend, Wierzbicka 1990) and their cognate adjectives are tested against such a benchmark, compared with lexicon that is not usually considered language-specific or showing a Russian «linguistic view of the world» (*prostranstvo* ‘space‘, *strast’* ‘passion‘). The English stems rather than words are analyzed as single items (e.g. *bore, boredom, boring* and *bored* are counted as one lexical item).

The Russian-English corpus and the Russian-Ukrainian corpus both were used for this analysis. Language-specificity is not a property of a given word of a given language *per se* but is defined in juxtaposition with another language or set or group of languages. Thus, a closely related language with a lot of cognate lexicon is chosen (Ukrainian) alongside with a far more distantly related English: effects may consequently vary.

The data are presented in separate tables for translation models (Table 1) and stimuli (Table 2).

	H (M)	Herf(M)	HerfNorm(M)	H (M) Ukr	Herf Ukr	HerfNorm (M) Ukr
poshlost’	0,800488	0,278025	0,261616	0,772807	0,210744	0,17316
udal’	2	0,25	0	0,758005	0,208889	0,152381
toska	1,0829	0,12835	0,115895	0,56419	0,4124	0,404561
prostranstvo	0,717253	0,337868	0,321719	0,394521	0,655047	0,649392
ujut	0,524006	0,317778	0,294253	0,327774	0,699074	0,690476
strast’	0,318889	0,730844	0,727207	0,317218	0,731852	0,728805
prostor	0,921035	0,132653	0,065934	0,163598	0,830579	0,822511

Table 1: Entropy, Herfindahl index and Normalized Herfindahl index for original Russian texts (translation patterns): Russian-English (columns 2-4) and Russian-Ukrainian (columns 5-8) corpora.

	H (S)	Herf(S)	HerfNorm(S)	H (S) Ukr	Herf (S) Ukr	HerfNorm (S) Ukr
poshlost’	1,186608	0,085432	0,064646	0,652852	0,265306	0,142857
udal’	3,429908	0,135802	0,049383	0,716003	0,2	0
toska	1,614182	0,038156	0,033298	0,677783	0,29896	0,293484
prostranstvo	0,634029	0,449467	0,440865	0,171447	0,861082	0,859953
ujut	0,928966	0,223081	0,203659	0,499347	0,550926	0,544601
strast’	0,991535	0,159184	0,134454	0,656611	0,439192	0,43341
prostor	1,344476	0,071834	0,061287	0,295317	0,780607	0,779531

Table 2: Entropy, Herfindahl index and Normalized Herfindahl index for translated Russian texts (translation stimuli): English-Russian (columns 2-4) and Ukrainian-Russian (columns 5-8) corpora.

4 Russian Candidates for Language-Specificity: General Discussion

It is found that for the word *udal’*, as compared to English, entropy is more than 3, it is found almost twice more frequently (instances per million) in original texts, and, of the total of 11 occurences, is translated in 9 different ways (*bravado, reckless jockeying, courage* etc.). It can be considered a typical «language-specific» word. At the same time, *toska* yields as many as 66 different English counterparts, of which *longing*, *yearning*, and *anguish* prevail. The entropy is lower than for *udal'*, but it is also more frequent in original texts.

The word *strast'* has the same frequency in a Russian text as *toska*, but it is clearly less language-specific as far as original texts are concerned: it is translated as *passion* in 85% cases and thus has low entropy. When we turn to the translations, however, we see that this word is used by different Russian translators as an umbrella term for a number of English words signifying different passions (such as *lust*, *obsession* etc.) and has consequently a higher "specificity" profile. This proves that different translation directions are to be treated separately.

Our method shows a clear-cut distinction between *prostor* and *prostranstvo* 'space', with the second being a rather routine translation (or original source) for English *space*, and the former being more language-specific, with a low monopoly index and high entropy. Two other words, usually considered language-specific (*ujut* and *poshlost*'), have less telling parameters. *Ujut* is, in fact, close to *strast'*: the English translators use either *coziness*, *snug* and *comfort* for it, whereas the Russian ones see in it an umbrella term for different feelings. A interesting example from Gogol shows all these three roots in translation:

(1) I kak chudna ona sama, èta doroga: yasnyj den', osennie list'ya, xolodnyj vozdux… pokrepche v dorozhnuju shinel', shapku na ushi, tesnej i *ujutnej* prizhmemsja k uglu! [N. V. Gogol'. Mertvye dushi (1835-1852)]

And how interesting for its own sake is a highway, should the day be a fine one (though chilly) in mellowing autumn, press closer your travelling cloak, and draw down your cap over your ears, and *snuggle cosily, comfortably* into a corner of the britchka. [Nikolay Gogol. Dead Souls (D.J. Hogarth, 1931)]

Poshlost', a word famously praised by Vladimir Nabokov (see above) as language-specific, has nevertheless a leading counterpart, *vulgarity*, but other variants are also widespread, and the translation entropy is rather high. (Note that the texts by Nabokov himself, added to the Russian-English/English-Russian parallel corpora in 2015, are not counted here. Both Nabokov as translator of his own prose and other people who translated it certainly paid particular attention to this lexical item which occurs frequently in the Russian versions of his texts).

Things change when we evaluate the language-specificness of these Russian words against Ukrainian. Both *prostor* and *prostranstvo* have a common Ukrainian cognate *prostir* and do not exhibit a cross-linguistic variation there. *Ujut* has also a clear counterpart, *zatyshok*, and only *udal'* and *poshlost'* qualify as language-specific even compared to a fellow Eastern Slavic language. *Toska* has a lot of Ukrainian counterparts (*tuha, sum, nud'ha* etc.), but there are still some clear leaders among translation equivalents.

5 Further Questions

The question rests open, however, whether all the properties and connotations of "language-specific lexicon" can be identified in such a way. For example, the Russian word *razluka* '~parting (of loving ones) and emotional experience connected with it almost always, in the Russian-French parallel corpus (including different translations of the same text), is translated *séparation*, clearly lacking nevertheless some connotations of the Russian word. At the same time, *razluka* is conditioned by different French stimuli – *éloignement, privation, quitter, absence*. This paradox shows that the stimuli can sometimes signal specificity more clearly than translation models.

On the other hand, it seems clear that the lack of a uniform translation at least not always corresponds to the problems of "linguistic view of the world" (as it is commonly understood by the authors using this term). For example, the German word *außerordentlich* lit. 'extraordinary (extraordinarily)', analyzed, for example, in (Добровольский 2013: 220-223), clearly qualifies as a language-specific word against the Russian equivalents. The Russian-German parallel corpus within the RNC (used also by Dobrovol'skij in the quoted paper) shows that it counts 53 lexically different models of translation in the German-Russian texts and is yielded by 30 different stumuli in the Russian-German texts. (Russian adjectives and adverbs with the same stem were counted as one item, given the fact that these categories are normally merged in German). The entropy is very high, more than for any word analyzed above (4,5 for models and 3,1 for stimuli), and the normalized Herfindahl index is 0,04 for models and 0,1 for stimuli, that is, closer to the best language-specific words like *toska* and *prostor*. (The entropy would be slightly lower if we count the Russian words with different suffixes *neobychno, neobychajno* and *neobyknovenno* as a single item, but even then still above 4,0 for models).

The differences between the models and the stimuli indices may be in this case be partly due to the diachronic factor which was not discussed above: the majority of the Russian original texts included into the Russian-German parallel corpus are classical texts of the 19th century when the Russian adjective *chrezvychajno* 'extraordinarily' was used more often than in modern texts and appeared as the main stimulus for the adverbial uses *außerordentlich*. Nevertheless even for these texts the indices are telling enough.

Can *außerordentlich* be considered to be a 'key element of the German *Weltbild*'? Even if we recognize that such a notion exists it would be more cautious to see here just an intensifier with a wider collocation range – and in this sense, specific for German. The Russian lexical units signifying something like 'very' are structured otherwise, requiring something like *udivitel'no* lit. 'astonishingly' for *dobryj* 'kind' to yield the sence (whereas *chrezvychajno dobryj* is obsolete; in Modern Russian it, if existed, would signify something like 'too kind', 'more kind than it is necessary'). Sometimes *außerordentlich* is not translated at all or rendered without a separate lexical counterpart (e.g. *außerordentlich geräumig* 'extraordinarily spacious' = *ogromnyj* 'enormous'). The abundance of synonymous terms signifying 'incredibly', 'extremely' etc. ("strong intensifiers") and their competition with a standard intensifier ('very') is well-known cross-lingustically although some predominant patterns do emerge (cf. Dahl 2004: 137, 139).

Polysemy may also be a key factor, as the German word in question (as well as *extraordinary* in English) has also the semantics of 'different from the ordinary', 'remarkable', 'unusual', 'outstanding', alongside with the meaning of intensifier 'extremely'; these different meanings multiply the choices for Russian translation counterparts. Cf. an example from an English-German parallel text:

(2) There was nothing so VERY remarkable in that; nor did Alice think it so VERY much *out of the way* to hear the Rabbit say to itself "Oh dear! Oh dear! I shall be too late!" [Lewis Carroll. Alice's Adventures in Wonderland]

Alice fand es auch nicht sehr *außerordentlich*, daß sie das Kaninchen sagen hörte, "O weh, o weh! Ich werde zu spät kommen!"

Perhaps such uses should be excluded during the statistical analysis, or, at least, they should be analyzed separately.

Thus, our study shows that language-specificity is rather a set of scalar values, it can be measured and tested on parallel corpora, and that things vary according to the direction of translation and language chosen for comparison and some other factors sketched above. If we study a set of Romance and Slavic languages, it would be possible to find whole clusters of language-specific lexicon for a language or a group of languages. An automatic extraction of language-specific lexicon in large corpora can be an interesting challenge.

The data drawn from such corpora can be used in bilingual dictionaries and in language teaching. It seems that a comprehensive list of possible translations is telling enough for the dictionary user, showing the difficulties of translation in different contexts. In this sense indeed every word is translatable, as the critics of the Neo-Humboldtianism correctly put it. The question is that this translation pattern may be more or less clearly defined, no matter whether it depends on culture and/or cognitive factors or not.

6 References

Добровольский, Д. О. (2009). Корпус параллельных текстов в исследовании культурно-специфичной лексики. In *Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы*. СПб.: Нестор-История, pp. 383—401.

Добровольский, Д. О. (2013). *Беседы о немецком слове*. М.: Языки славянских культур.

Павлова, А. В. (сост.) *От лингвистики к мифу: лингвистическая культурология в поисках «этнической ментальности»*. СПб.: Антология, 2013.

Руссо, М. М. Неогумбольдтианская лингвистика и рамки «языковой картины мира». In Политическая лингвистика, 1 (47), pp. 12—24. Accessed at: http://journals.uspu.ru/attachments/article/622/%D0%9F%D0%BE%D0%BB%D0%B8%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_2014_1_%D1%81%D1%82.%2001.pdf [25/06/2016]

Зализняк, Анна А. & Левонтина, И. Б. & Шмелев А. Д. (2012). *Константы и переменные русской языковой картины мира*. М.: Языки славянских культур.

Cysouw, M., & Wälchli B. (eds.). Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in *Sprachtypologie & Universalienforschung STUF*, 60.2.

Dahl, Östen (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins.

Boroditsky, L. (2011). How Languages Construct Time. In Dehaene and Brannon, E. (eds.) *Space, time and number in the brain: Searching for the foundations of mathematical thought*. Amsterdam: Elsevier.

Levinson, S. C. (2003). *Space in language and cognition: explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Lucy, J. A. (1992), *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge: Cambridge University Press.

Sériot, P. (2005) Oxymore ou malentendu? Le relativisme universaliste de la métalangue sémantique naturelle universelle d'Anna Wierzbicka. In *Cahiers Ferdinand de Saussure*, № 57.

Wierzbicka, A. (1990) Duša ('soul'), toska ('yearning'), sud'ba ('fate'): three key concepts in Russian language and Russian culture. In Zygmunt Saloni (ed.). *Metody formalne w opisie* *języków słowiańskich*. Białystok: Białystok University Press, p. 13–36.

Acknowledgements

I thank Dmitri Dobrovol'skij, Anna Zalizniak and Vladimir Plungian for some comments and material for this article.