Sense Ranking in Dual-Language Online Dictionaries

Mark Kit[1], Elena Berg[2]

Abstract

One of the most essential components in the foreign language learning process are online dual-language dictionaries. The efficiency in obtaining required lexical data from them directly translates into the efficiency of the learning process. Among other aspects, the order of presentation of retrieved translations is of great importance. Developers of online dictionaries often merge different source dictionaries in a single dataset, which results in translations placed in non-systematic order.

Ranking using word frequencies obtained from national corpora does not solve the problem since the place of the translation in the list depends not from how often the word occurs in texts, but rather from how frequent the sense in which this word is used occurs. Besides, ranking must be done separately for translations belonging to different part-of-speech groups (nouns, verbs etc.). To understand the magnitude of the ranking effort, the scope of work needed should be determined beforehand.

This paper describes the analysis of lexical units to be ranked in the lexical dataset used in the online dictionary LexSite. A subset made for this analysis includes 45,625 English words and 24,628 Russian words with corpora-based ranks below 60,000 in 4 part-of-speech categories (nouns, verbs, adjectives and adverbs). The study found that words with large number of lexical senses makes up about 3% of all words in the subset. Although words with small number of senses are in abundance (e.g. 3,955 English words in the 5-10 senses category), large efforts are unnecessary for these groups because small lists of translations normally are easy to understand.

The paper discusses detailed results of the studies described above and substantiates the need and techniques for translation ranking in online dual-language dictionaries.

Keywords: lexicography, word frequency, ranking, online dictionary

1. Introduction

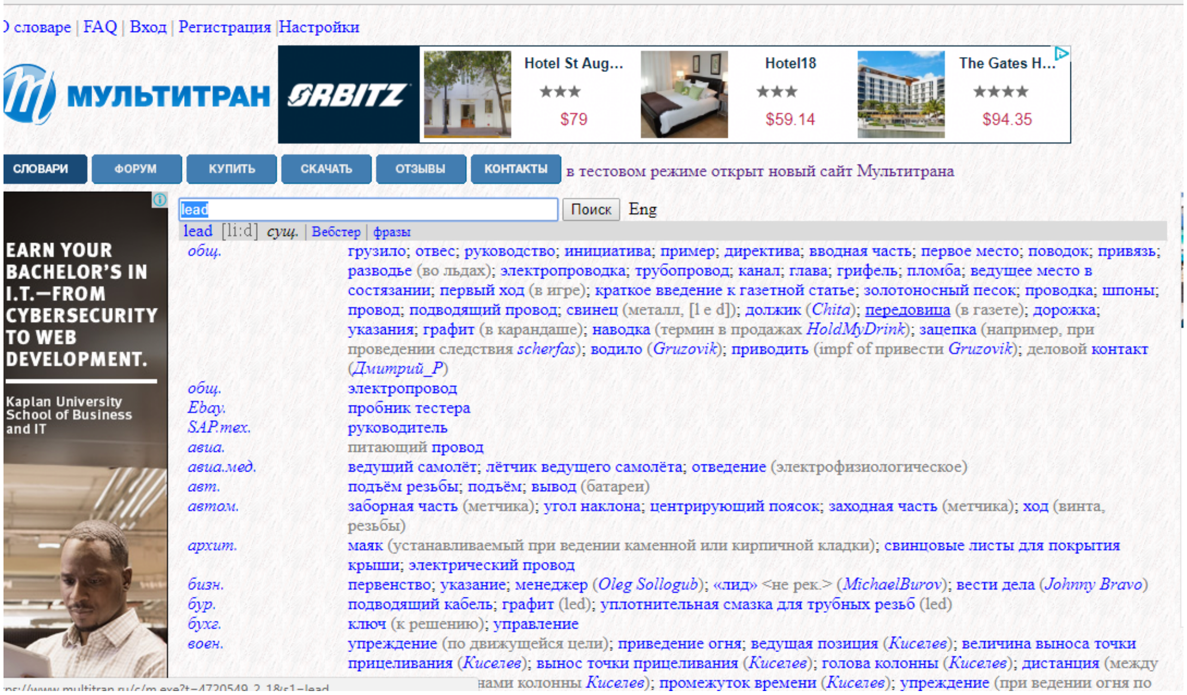
Online dual-language dictionaries are among the most essential means employed in foreign language studies. The efficiency in obtaining required lexical data from these resources directly translates into the efficiency of the learning process. Despite of endless opportunities offered by new technologies, researchers point out that creation of electronic/online dictionaries face many challenges [1].

Among other aspects, the order of presentation of retrieved translations is of great importance. Developers of online dictionaries often merge different source dictionaries in a single dataset, which results in translations placed in non-systematic order.

Since the developers of online dictionaries tend to expand their coverage as much as they can, they often merge different dictionaries into a single large dataset, mixing together special, professional and common usage terms and expressions. As a result, in response to the user's query such a system produces a large number of translation, often placed in a random order. Below is an example of translations of the word *lead* obtained from a popular dictionary Multitran (Fig. 1). There are about 1,500 translations among which the user has to find the one most relevant to the specific discourse situation.

[1] Language Interface Inc. (USA)

[2] Ural State Law University (Russia), Language Interface Inc. (USA)


) словаре | FAQ | Вход | Регистрация |Настройки
МУЛЬТИТРАН
ORBITZ
Hotel St Aug...
★★★
$79
Hotel18
★★★
$59.14
The Gates H...
★★★★
$94.35
СЛОВАРИ
ФОРУМ
КУПИТЬ
СКАЧАТЬ
ОТЗЫВЫ
КОНТАКТЫ
EARN YOUR BACHELOR'S IN I.T.—FROM CYBERSECURITY TO WEB DEVELOPMENT.
Kaplan University School of Business and IT
lead
Поиск
Eng

в тестовом режиме открыт новый сайт Мультитрана

lead [li:d] *сущ.* | Вебстер | фразы

общ.	грузило; отвес; руководство; инициатива; пример; директива; вводная часть; первое место; поводок; привязь; разводье (во льдах); электропроводка; трубопровод; канал; глава; грифель; пломба; ведущее место в состязании; первый ход (в игре); краткое введение к газетной статье; золотоносный песок; проводка; шпоны; провод; подводящий провод; свинец (металл, [l e d]); должик (*Chita*); передовица (в газете); дорожка; указания; графит (в карандаше); наводка (термин в продажах *HoldMyDrink*); зацепка (например, при проведении следствия *scherfas*); водило (*Gruzovik*); приводить (impf of привести *Gruzovik*); деловой контакт (*Дмитрий_P*)
общ.	электропровод
Ebay.	пробник тестера
SAP.тех.	руководитель
авиа.	питающий провод
авиа.мед.	ведущий самолёт; лётчик ведущего самолёта; отведение (электрофизиологическое)
авт.	подъём резьбы; подъём; вывод (батареи)
автом.	заборная часть (метчика); угол наклона; центрирующий поясок; заходная часть (метчика); ход (винта, резьбы)
архит.	маяк (устанавливаемый при ведении каменной или кирпичной кладки); свинцовые листы для покрытия крыши; электрический провод
бизн.	первенство; указание; менеджер (*Oleg Sollogub*); «лид» <не рек.> (*MichaelBurov*); вести дела (*Johnny Bravo*)
бур.	подводящий кабель; графит (led); уплотнительная смазка для трубных резьб (led)
бухг.	ключ (к решению); управление
воен.	упреждение (по движущейся цели); приведение огня; ведущая позиция (*Киселев*); величина выноса точки прицеливания (*Киселев*); вынос точки прицеливания (*Киселев*); голова колонны (*Киселев*); дистанция (между нами колонны *Киселев*); промежуток времени (*Киселев*); упреждение (при ведении огня по

Fig. 1. Example of unordered translations presented by online dictionary Multitran.

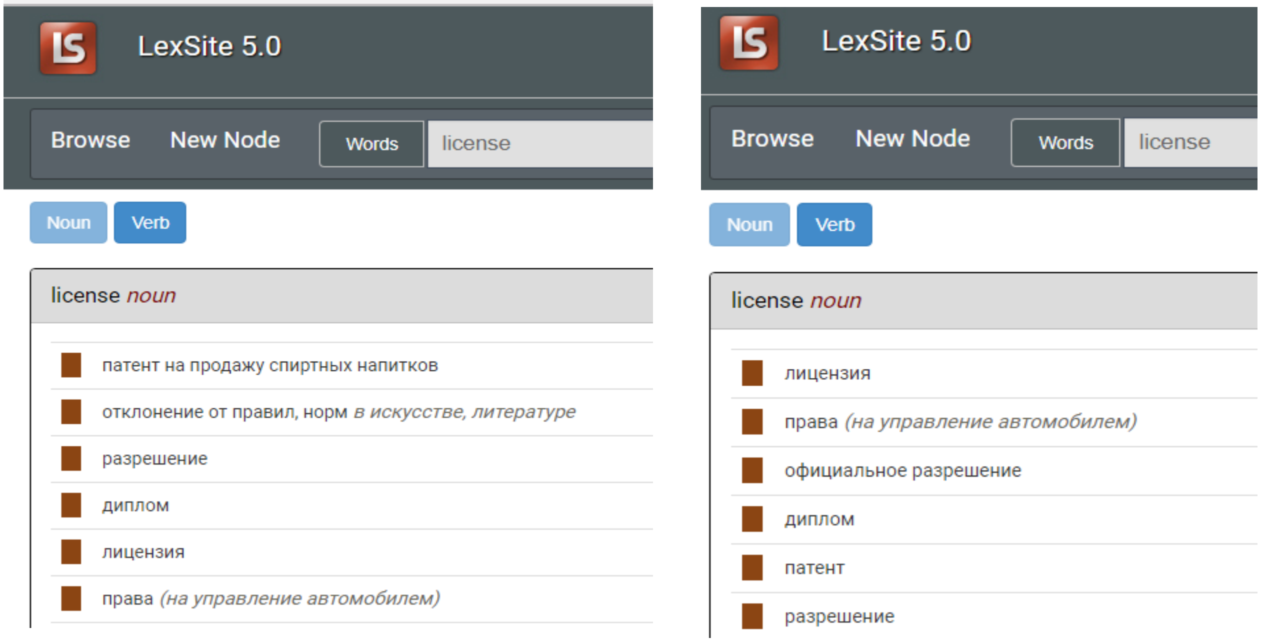
The authors of traditional printed dictionaries streamline the search for the right translation by placing the translations in the decreasing order of their usage frequency. When migrating these dictionaries to the electronic databases and merging different dictionaries, order of translations changes and the search becomes labor-intensive.

The rank of the lexical unit is the inverse of the number word's place in the usage frequency list. For instance, the English word *the* has rank 1 since it is the most frequent word found in the English corpora, the word *be* has rank 2 because its occurrence frequency is next to that of word *the*. These ranks are extracted from large corpora and, seemingly, can be used to automatically order lexical units in the dictionaries. However, this approach does not help when it comes to dual-language dictionaries, because the translations should be ordered based on the rank of the meanings rather than the rank of the word.

For example, the Russian word *собака* has a rank of 434 while the word *парень* rank is 369. Ordering translations by word usage rank would put the word *парень* on the top of the list while the word *собака* that most commonly used for *dog* will be placed below.

The reasons stated above demonstrate that ranking of the translations should be sense-based. In this approach, ranks are assigned to the source-target pairs for each set of translations belonging to any particular source word. It should be noted that for translations to the opposite direction the ranking will be different since the sets of translations source-target and target-source are different. Besides, ranking must be done separately for translations belonging to different part-of-speech groups because the frequencies of usage of translations for, e.g. word *split* as a noun has nothing to do with frequencies of translations of word *split* as a verb.

The screenshots below show the order of translations before and after ranking conducted using the online lexical-semantic platform LexSite [2] developed in the language engineering company Language Interface. It can be seen that raking of the word *license* was done separately for nouns and verbs so that in the noun group the word *лицензия* is on the top of the list. Similarly, in the verb group the word *лицензировать* moves to the top of the list.



license *noun*

- патент на продажу спиртных напитков
- отклонение от правил, норм *в искусстве, литературе*
- разрешение
- диплом
- лицензия
- права *(на управление автомобилем)*

license *noun*

- лицензия
- права *(на управление автомобилем)*
- официальное разрешение
- диплом
- патент
- разрешение

Fig. 2. Order of translations of the word *license* before and after sense-based ranking, noun group.

The magnitude of ranking effort can be quite significant, thus posing an obstacle for the dictionary developers. “In principle, of course, these difficulties could be resolved if enough time and, consequently, money were to be devoted to editing the preliminary version of the dictionary...” [3]. The recognition of this difficulty necessitates the need for planning of the ranking work and evaluation of its scope.

2. Discussion

There are many approaches to sense ranking, including chronological, markedness, frequency and logic [4]. We suggest that the best approach should be based on “the relevancy of the term sought to the context of previous searches.” [5]. Yet the initial ordering of senses (when the dictionary has not yet acquired the knowledge of the user’s context) shall be based on sense frequency.

The analysis of lexical units to be ranked was run using the lexical dataset of the above-mentioned online Russian-English lexicographic platform LexSite. The subset extracted from that dataset includes 35,145 English words and 23,892 Russian words with corpora-based ranks below 60,000 in 4 part-of-speech categories (nouns, verbs, adjectives and adverbs). Data on the content of the subset subjected to the analysis is shown in Table 1.

Table 1. Content of the lexical data subset used in the experiments.

	English	**Russian**
Nouns	19,992	12,063
Verbs	4,958	5,700
Adjectives	8,919	5,004
Adverbs	1,275	1,124
Total	35,145	23,892

The paper describes results of the studies described above and substantiates the need and techniques for ranking in dual-language online dictionaries.

The following strategy is proposed for the planning of ranking efforts:

- Ranking is not required where a POS group includes a single translation of a source word

- Where a POS group includes a small number of translations (1-5) ranking is unimportant since it is easy to find the needed translation among the offered ones regardless of their order
- Words with greatest corpora-based frequencies should be ranked first they are in the greatest demand.

This study shows that vast majority of English words in the subset has low polysemy. In fact, the polysemy of 78% of these words is less or equal to 10. **Error! Reference source not found.** shows the polysemy distribution of English words in the subset.

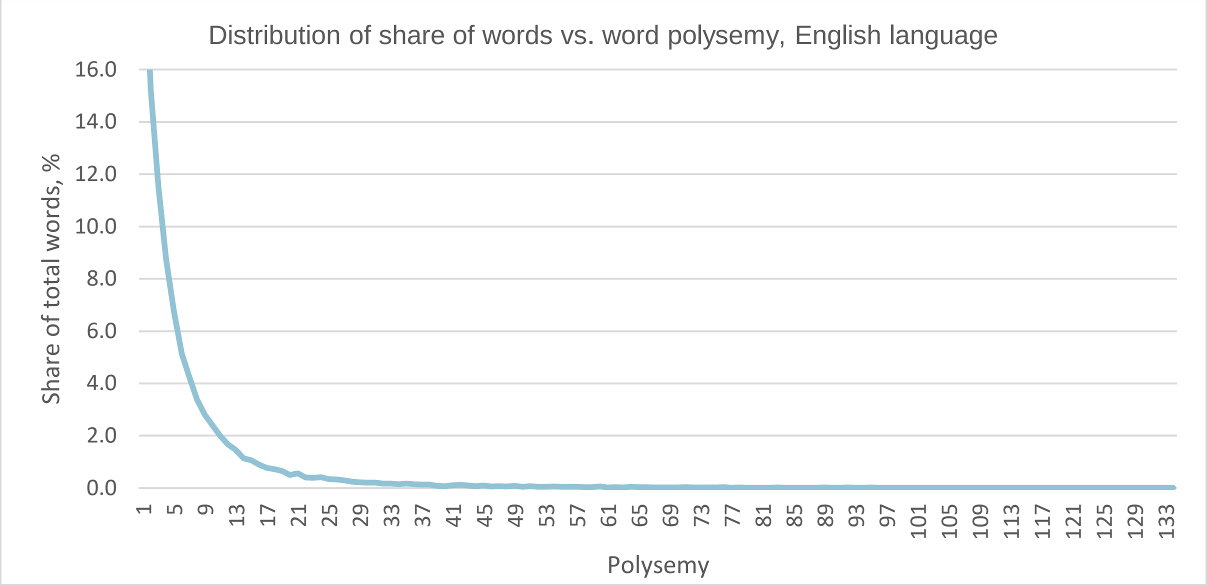


Fig. 3. Distribution of words with different polysemy in the English language.

Similar distribution is observed in the Russian subset. More than 80% of all words of the subset have polysemy less or equal to 10. The polysemy distribution for the Russian language is shown on Fig. 4.

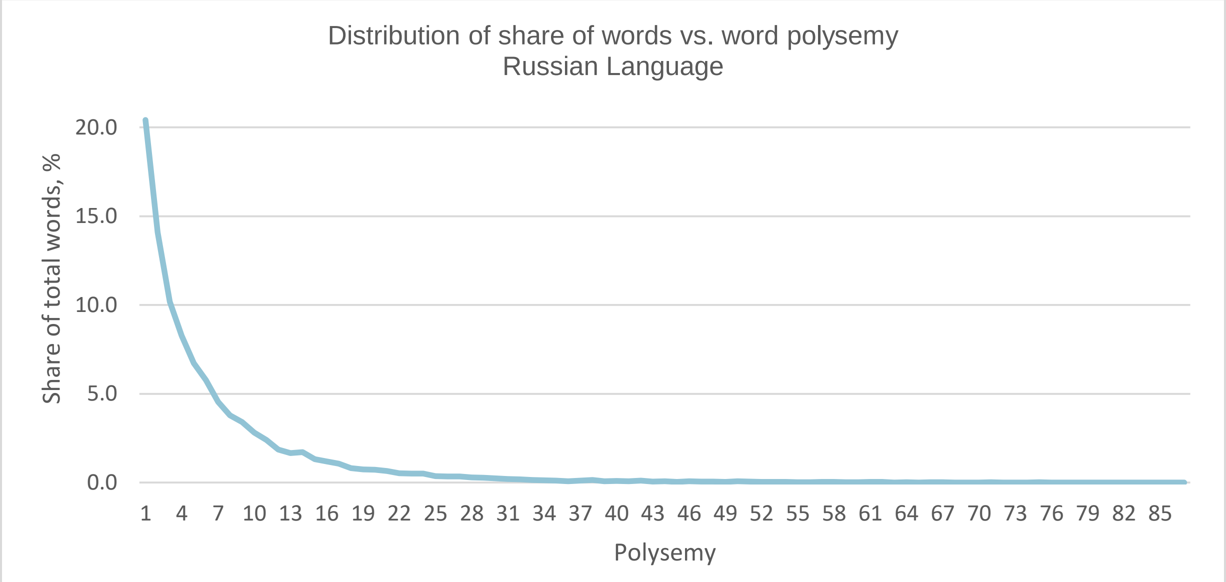


Fig. 4. Distribution of words with different polysemy in the Russian language.

The study confirmed the assumption that words with large number of lexical senses comprise a small portion of all words in the subset. Although words with small number of senses are in abundance, their ranking may not be required because small lists of translations normally are easy to understand.

No less important is the correlation between polysemy and corpora-based ranks. The analysis showed that the average polysemy drops as the corpora-based rank increases. This means that, generally, words used most frequently have greater polysemy and sense-based ranking of these words will take more efforts. Fig. 5 shows how many words of certain polysemy fall into given rank ranges.

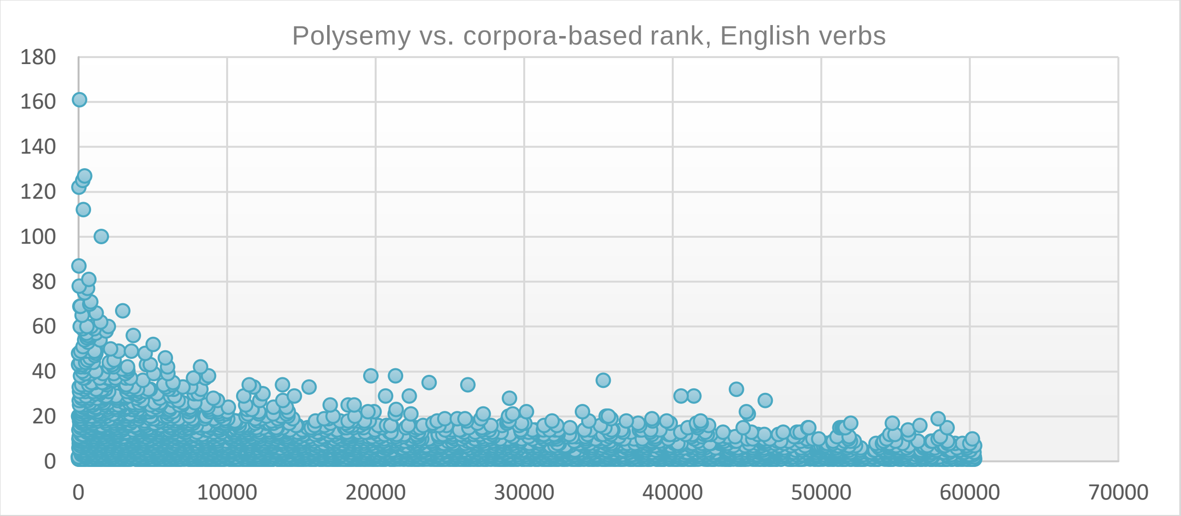


Fig. 5. Distribution of polysemy vs. corpora-based rank for English verbs.

3. Conclusion

This study can form a basis for work planning when making dual-language dictionaries that would benefit a broad range of language learning assignments. Besides, the results of this study will help developing sense-ranking assignments for students studying foreign languages so that they would improve their lexical proficiency through ranking words on their own. The language teachers, on the other hand, can use this data when making learning dictionaries focused on specific pedagogical needs.

References

[1] De Schryver, G.-M. “Lexicographers' Dreams in the Electronic-Dictionary Age”, International Journal of Lexicography, Oxford Academic, 2003, Vol. 16, Issue 2, p. 143–199

[2] http://www.lexsite-dictionary.com

[3] Adamska-Sałaciak, A. “Issues in compiling bilingual dictionaries”, http://www.academia.edu/5198166/Issues_in_compiling_bilingual_dictionaries

[4] Lew, R. “Identifying, ordering and defining senses”. http://www.staff.amu.edu.pl/~rlew/pub/Lew_2013_Identifying_ordering_defining_senses_[preprint].pdf

[5] Kit, M., Kit, D. “On Development of “Smart” Dictionaries”, Cognitive Studies, Warsaw, SOW Publishing House, 2012, p. 115-127