COGNITIVE STUDIES | ÉTUDES COGNITIVES, 12
SOW Publishing House, Warsaw 2012

**MARK KIT**[1], **DMITRY KIT**[2]
[1]Language Interface Inc., New York
mark.kit@langint.com
[2]University of Texas, Department of Computer Science, Center for Perceptual Systems, Austin
dkit@cs.utexas.edu

# ON DEVELOPMENT OF "SMART" DICTIONARIES

**Abstract**

The paper discusses the need for development of intelligent dictionaries that allow for two-way interaction with its users. Theoretical ground for such development is suggested. Practical implementation as LexSite lexical resource is shown, concepts for further improvement of the efficiency are proposed.
**Keywords**: lexicography, online dictionaries, translation technology, smart dictionary, intelligent dictionary, semantic representation, lexical units.

## 1. Introduction

Every hour at least 130,000 pages of texts are being translated in the world[1]. The vast majority of these texts are of informational/technological nature and they should be translated very fast and very accurately. In these conditions the efficiency of translators' tools, first and foremost – online dictionaries – is of critical importance. Characteristics of online dictionaries make a considerable impact on translator's performance and quality of his or her translation.

Undoubtedly, online dictionaries are very helpful when it comes to translation. The translator can instantly receive the sought after results. The lexicographers can update the dictionary any time and from any place. These dictionaries are available to anyone from anywhere, using PCs, notebooks, smartphones, and electronic tablets. It may look like the lexical support of translators is no longer a problem. However, this is not the case.

To understand the problem one has to view it from the translator's point of view. Our analysis shows that when working on a translation the translator on average makes 45 dictionary queries per hour (the range found in our experiment was from 17 to 63 dictionary calls an hour). At this frequency time spent waiting for dictionary response or on searching for the appropriate translation in the output data results in tremendous loss of performance (up to 74%) (Kit 2010: 151).

[1]This figure is obtained from the total market size (Nataly Kelly, Robert Stewart, 2010), the average translation rate and the ratio between freelance and corporate translators (EUATC, 2006).

Further performance is also lost because the translator loses focus on the text being translated. The longer the attention is taken away from the text, the longer it will take to get back to the work again.

One has to recognize that the user calls the dictionary with the purpose to receive the only translation that is best for the text he works on. In the ideal world the dictionary would give him that result and nothing else. The current dictionaries, however, return a flow of information, which may even not contain the required translation. For example, one of the most popular English-Russian online dictionaries in response to the query "barrel" (meaning "gun barrel" in the text) returned more than 400 words where only 4 words contained useful information. Another dictionary (most popular) produced a page containing 940 words where the required translation was composed of 2 words. Thus, the useful content in the information received was 1% and 0.2%, respectively. This puts a heavy burden of finding the required result on the translator's shoulders.

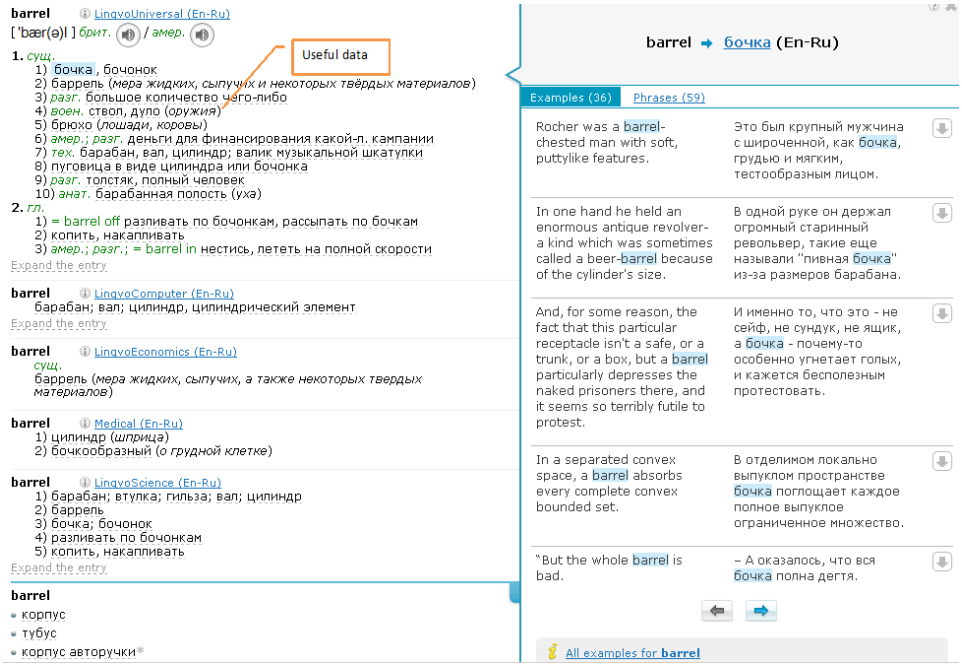


**Figure 1.** A screenshot of a page returned by a popular English-Russian dictionary in response to the word "barrel". If the user is searching for the meaning "gun barrel" then these results only contain 0.2% useful content.

A typical timeline of a single dictionary interaction is illustrated in Figure 2. Durations of each step are shown as an example, but represent typical values. This diagram shows the most favorable case where the dictionary returns the required translation (among other results) in the first search, but in the alternative case the total duration of a dictionary interaction is even greater.

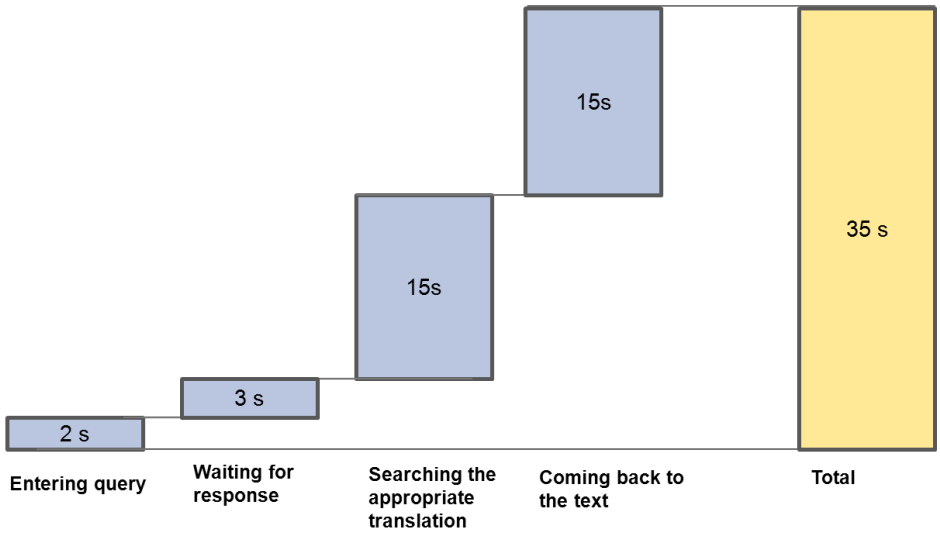


**Figure 2**. Example of a dictionary interaction timeline which starts when the user entered a query and ends with the user getting back to the current translation (blue rectangles). The time inside the rectangle is the duration of each stage. The yellow rectangle is the total time of the interaction.

As an example of how long it takes to comprehend the text to be translated, the phrase below was taken from an ordinary document, one of many a translator encounters daily.

> "*The Buyer agrees with the Seller that where the Buyer or the Company is paid any amount or receives any value in respect of a Refinery Claim it shall pay, or procure the payment of, to the Seller the amount or value so received less any reasonable costs incurred by the Buyer in obtaining such amount (to the extent that such costs have not been reimbursed pursuant to the indemnity in paragraph 19.6) less any Tax suffered by the Buyer or the Company on such receipt within five Business Days of receipt of such amount provided that clauses 9.6(b) and 9.7 of this Agreement shall not apply to such payment.*"

This is a typical sentence taken from a legal document. The document was 68 pages long and its translation was scheduled to be delivered in just 36 hours after submission of the original document. If the translator gets distracted even for a few minutes to make a search in the dictionary, his concentration on the text, and therefore time, is lost.

All this suggests that the traditional view of online dictionaries such as "same as those printed on paper but implemented with electrons" has reached its limits and needs to be revised. To check this conclusion and initiate development of efficient translation dictionaries, Language Interface Inc. (USA) opened the Experimental Platform LexSite project. The primary purpose of the project is to create a platform

where improved online dictionaries can be developed and built. The project includes a suite of English-Russian-English dictionaries as its core lexical resource.

In late 2009 LexSite was made publically available on the Internet. Later it was updated and modified; this is a continuous improvement process. To improve responsiveness, its design employs Web 2.0 technology. The page also maximizes the essential information provided to the user, with additional data presented as needed. For example, synonyms are provided upon user's request.

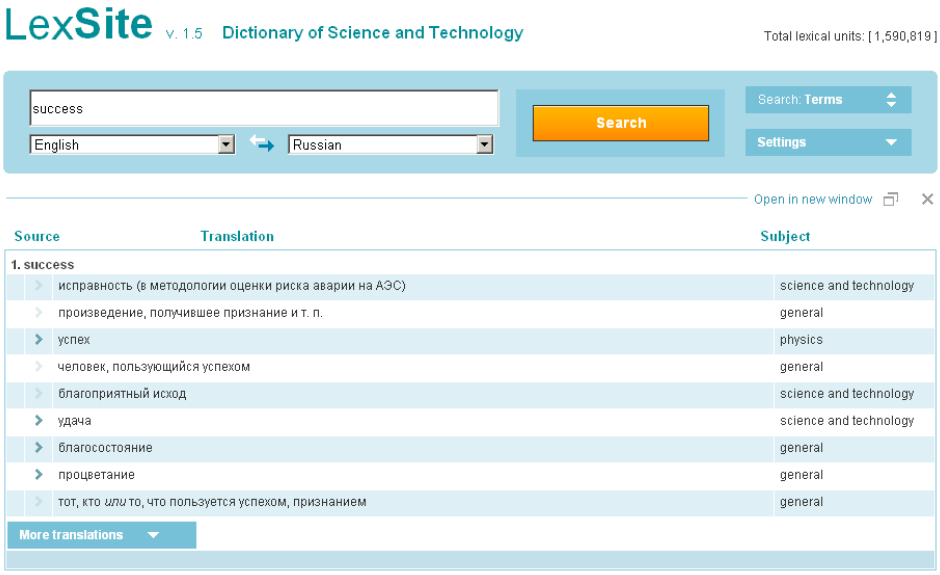


**Figure 3**. Lexical resource LexSite. The results are minimal, which facilitate quick searching by the translator. The dark arrows on the left can be clicked to obtain synonyms for that particular meaning.

The dictionary allows the user to enable any combination of subject filters. Furthermore, the users can search examples of usage of lexical units in the parallel corpus included in the resource.

Today the dictionary ensures fast response (a stress-test resulted in 400ms under a load of 36,000 queries per hour). The next objective of the developers is to improve the relevance of the output results. The system is scalable and can be quickly extended as needed.

Thorough tune-up of the translation search mechanism ensures that the user receives the most relevant values. For example, in response to query "broke in" the user is offered foreign equivalents of "break in", "break-in" and "broken-in". This makes the interaction between the user and the dictionary very short. However, the developers believe that this is just a starting point for creating smart dictionaries.

**2. Smart dictionary**

Further improvement of efficiency consists of prioritizing translations to be displayed to the user. If the dictionary finds more than one meaning of the lexical unit sought, what should be displayed at the top of the list?

One way to answer this question is based on what type of user the dictionary is working with. For professional translators or scientists working with a complicated text, the least frequent words may become more important than the common ones, because this user knows the language well. Ignoring this type of user and employing the word frequency curve to prioritize results may not be the best strategy in this case.

Another approach for prioritizing results could rely upon the relevancy of the term sought to the context of previous searches. For example, if the user is asking for the word "well" that has been preceded by searches of the word "casing" and the term "blowout preventer" we have a good deal of confidence that the word "well" means "a deep hole or shaft sunk into the earth to obtain water, oil or gas". As another example, by itself the word "relief" can be anything, but if the text deals with high pressure vessels, then most likely it means "pressure relief".

To solve this problem the information obtained across multiple dictionary calls can be analyzed in order to study users' behavior, search patterns in their queries could determine the relevancy of the current search results.

The problem has two dimensions - semantic and temporal. The semantic dimension can be used to detect meaning of requested data while the temporal dimension enables us to make inferences on the character of the text, such as the field of knowledge it relates to or its complexity.

In the temporal dimension we deal with single calls to dictionary that, collectively, make up a multi-tier structure of dictionary calls (Figure 4).

- *Query* is a single search in the dictionary made by the user
- Queries make up *sessions* that last from one long break to the next one
- Collection of sessions are called *cycles,* which consist of all queries ever made by the same user
- Collection of cycles makes up a search *corpus.*

The multi-tier structure of dictionary calls shows that different granularity can be selected when analyzing dictionary interactions. The corpus detects patterns in the entire set of queries. A cycle tells us something about characteristic features of the texts being translated by a specific user, about this user's pace of work and personal patterns, such as durations of their sessions or frequency of calls. The session suggests data on the lexical composition of a specific text.

Temporal analysis sets the stage for initial strategies that should be utilized when servicing this particular user. The system can determine whether the user is a professional translator or an amateur whose knowledge of source/target language is quite poor. Random sessions, which are highly variable in terms of duration, would suggest that the user is, most likely, a specialist (an engineer, researcher or

physician) who has to work with foreign literature every now and then. Regular long sessions indicate that the user is a professional translator.

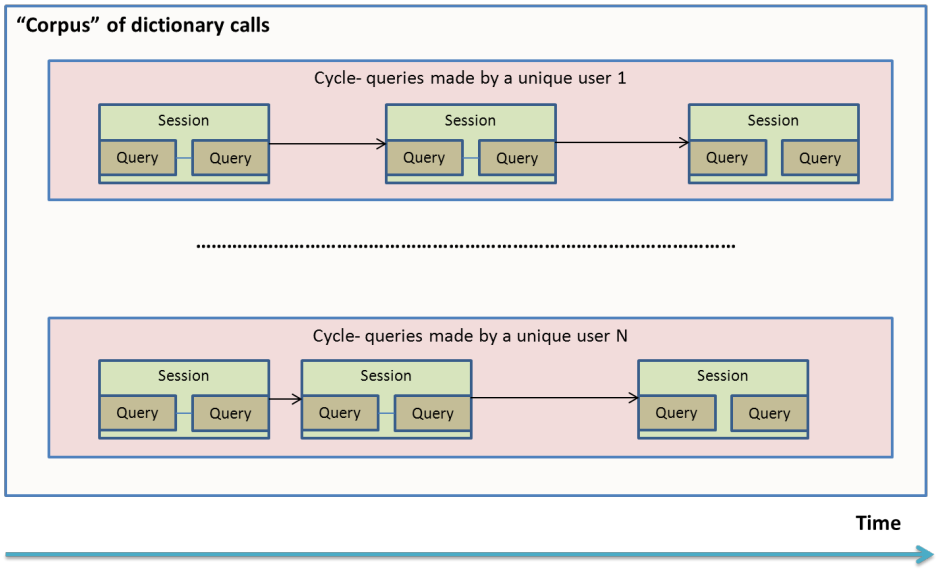


**Figure 4**. Multi-tier structure of dictionary calls. A collection of queries gathered over a short period of time are called *sessions* (green boxes). A *cycle* is a collection of all sessions made by a single user (pink rectangles). A *corpus* is a collection of all cycles across all users.

## 3. Traveling through the user's space

Even though we provide the user with highly relevant (HR) responses, leaving less relevant below fold (that can be shown too by clicking the "More" button), by doing so we achieve only the initial filtering of the output.

To clean up the search results from irrelevant lexical units the dictionary should "know" something about the user and the text being translated. Let's call it "user's space". Initially we do not know anything at all about the user's space and the distribution of user's queries is of uniform nature, i.e. all potential entries have equal probabilities. But as soon as the user starts making dictionary calls our perception of his space changes. In a few calls we can figure out what kind of user he is.The traditional way a user interacts with a dictionary is shown in Figure 5. The user enters a sequence of lexical units $u^s{}_i$ belonging to some source language $L_s(u^s_1, u^s_1 ... u^s_1 \in L_s)$ and in response receives sets of target language, $L_t$, equivalents $u^t_i$ of these units $(u^t_1, u^t_1 ... u^t_1 \in L_t)$. The set of resulting lexical units belonging to $L_t$ can be of any cardinality as shown in Table 1.

This approach is based on purely semiotic relationships between lexical units that belong to different languages. It is known that a source lexical unit (e.g., Russian) that is represented as $u^s_i$ can be represented as one of units $u^t_i$ in another language (e.g. English). This approach does not involve the semantic aspect at

all. In fact, what the user is looking for is a representation (in the form of a target language word or a combination of such words) of the meaning, not words.

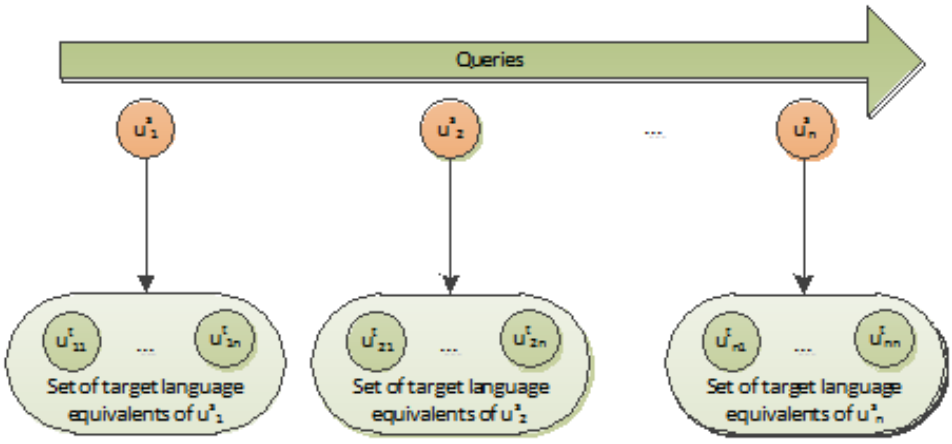


**Figure 5**. User's interactions with a dictionary. $u_i^s$ represents a lexical unit in the language $L_s$, $u_i^t$ — lexical units in the language $L_t$.

**Table 1**. Potential sets of target language units produced in response to user calls

| Case | Target set cardinality | Description |
|---|---|---|
| 1 | C=0 | No equivalents of the source unit found in the target language (to the extent it is represented in the dictionary) |
| 2 | C=1 | One-to-one match between the source and the target lexical units |
| 3 | $0>C>\infty$ | Most cases, which are due to homonymy |
| 4 | $C=\infty$ | Purely theoretical case or representation of a nonsense that can be anything |

The meaning can be determined indirectly, based on the user's previous searches and the initial conditions (priors). Initially it is not known what the priors are and are assumed to be uniform, with each possible meaning having equal probability of being correct.

These priors can then be updated (e.g. Bayes' rule) based on the collection of previous sessions (see Figure 4) – the user cycle. Moreover, even if a new user is encountered, it can be initially assumed that this user belongs to the most probable category of users and deals with the most probable category of texts, as determined based on the "corpus" of users' calls. Even such a simple analysis of the priors makes a great deal of difference. Table 3 shows examples a set of the kind of recommendations that can be made using the results of this analysis.

**Table 2**. The priors

| | |
|---|---|
| Type of user | Limited vocabulary, poor language skills |
| | Skilled language user |
| Type of text | General purpose text |
| | Special text (e.g. medicine, control systems) |
| Domain | General purpose |
| | Chemistry |
| | Medicine |
| | etc. |

**Table 3**. Dictionary settings based on the priors

| **Type of user** | **Type of text** | **Part of Zipf curve the user is working on** | **Features** |
|---|---|---|---|
| Limited vocabulary, poor language skills | General purpose text | Head | Spelling corrector<br><br>Lemma prompts (e.g. 'broke' is the past tense of 'break') |
| Skilled language user | Special text (e.g. medicine, control systems) | Tail | Reduced lemmalization (e.g., 'broke' is not used as a past of 'break') |
| Limited vocabulary, poor language skills | Special text (e.g. medicine, control systems) | Tail | Lemmalization, extensive spelling corrections |
| Skilled language user | General purpose text | Head | Synonyms |

Having calculated the priors can be used to start analyzing the current and future user sessions. One way to see the user's interaction with a dictionary is to focus on the domain (e.g. chemical industry related). Given a temporal sequence of search units $u_1$,$u_2$,... $u_t$, the most probable domain label $L$ can be determined as the L that maximizes equation 1.

$$\mathrm{P(L \mid u_{1:}}_t) \tag{1}$$

Due to homonymy the lexical units sought can belong to multiple domains. The most valuable information will be derived from those units that belong to a single domain.

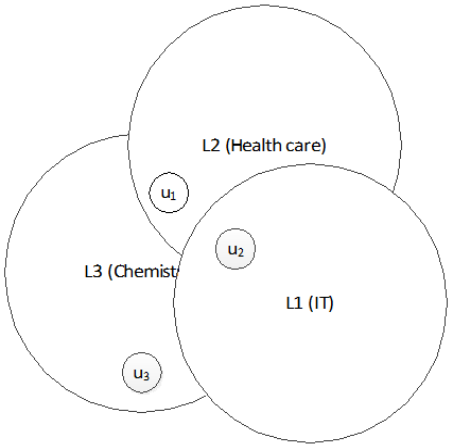


**Figure 6**. Lexical units distributed across domains. More common units ($u_2$) provide no information, while less common units ($u_3$) provide the most.

In Figure 6 L1, L2, and L3 represent the collection of words found in documents belonging to domains those labels are assigned to. The user starts a session by searching for the term $u_1$. This term could belong to domains labeled L2 or L3. The term $u_2$ provides no further information as the most likely candidates remain the same. Only after the search $u_3$ the current session can be classified as belonging to domain L3. Thus, the most valuable cases are those user calls that produce sets of target words with cardinality of 1 (see Table 1). A term belongs to more than one domain a judgment should be made to determine what label shall be used for the word (prioritizing the results based on Equation 1).

Suppose the word "headroom" is encountered. It can refer to a variety of labels $L_1$,$L_2 \ldots L_{n.}$. To determine which label $L$ is the most relevant in the current situation we need to know the distribution of probabilities over the entire variety for the given lexical unit U, (P($L_i/U$)). This distribution can be different for different labels.

Knowing that the user is currently dealing with telecommunication-related domain, the dictionary could select the most probable label $L_5$ from the set of potential labels for the given word.

Equation 1 may or may not use the temporal information of search terms. If time is ignored then the probability distribution may be approximated by:

$$\mathrm{P}(\mathrm{L} \mid \mathrm{u}_{1:t} = \frac{\Sigma_{i=0}^{T} \frac{u_i^L}{W_L}}{\Sigma_{j \in L} \Sigma_{i=0}^{T} \frac{u_i^j}{W_j}}), \tag{2}$$

where $u_i^L$ is the number of times the term $u_i$ appears in the documents labeled $L$ and $W_L$ is the number of words that appear in documents labeled $L$.

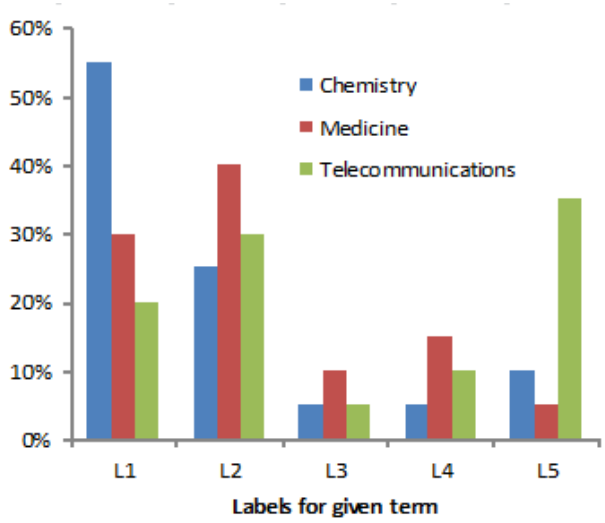


**Figure 7**. Example of distributions of labels for some word in various domains.

## 4. Implementation

An actual example of the history of queries made by a user in one session is shown in the table below (Table 4). Here the words that belong to one or very few domains shown in shaded boxes.

**Table 4**. An example history of user queries as recorded by LexSite. Underlined boxes mark the terms that appear in very few domains.

| | Output count | Comment |
|---|---|---|
| behind | 9 | |
| pancreas | 1 | |
| epigastrium | 3 | |
| bile | 4 | |
| participate | 6 | |
| beneath | 6 | |
| serve | 52 | |
| container | 24 | |
| extent | 23 | |
| retort | 26 | |
| chief | 23 | |
| undergo | 5 | |

| | | |
|---|---|---|
| underdog | 6 | |
| further | 19 | |
| estimate | 35 | |
| authorities | 3 | |
| minor | 29 | |
| birth | 14 | |
| interactions | 7 | |
| extremely | 6 | |
| carries | 21 | |
| carrie | 0 | Misspelled |
| islet | 2 | |
| susceptible | 11 | |
| excitation | 21 | |
| recognize | 24 | |
| relate | 15 | |
| content | 58 | |
| content | 58 | |
| intracellular | 3 | |
| formatoin | 0 | Misspelled |
| formation | 57 | |
| occur | 16 | |
| remains | 10 | |
| remains | 10 | |
| slightly | 6 | |
| heredity | 4 | |
| evolution | 21 | |
| development | 80 | |
| consider | 22 | |
| requir | 0 | Misspelled |
| requirs | 0 | Misspelled |
| require | 10 | |
| enzyme | 5 | |

| | | |
|---|---|---|
| suspect | 16 | |

From this example history a few inferences can be readily made. First, apparently the user does not know English very well; otherwise he would not search for the word "behind". Misspelled word "require" also supports this hypothesis, as well as other common words (extent, suspect, occur). Second, the user is dealing with a text related to medicine or health care ("pancreas", "epigastrium"). At this point the dictionary can start autonomously configuring its filters.

Table 5 shows an excerpt from a search history of another user. Here the user is most certainly a highly skilled translator since the dictionary is only searched for terms and almost never for common words. Inferences can be done instantly as the text is undoubtedly about some industrial chemical processes ('catalyst bed' suggests that, as well as 'ammonia synthesis'). This prompts to move chemistry-related terms to the top of the list while the commonly used words should receive much lower priority.

**Table 5**. Another example of a search history made by a different user (as the one in Table 4).

| |
|---|
| three-bed |
| three-bed basket |
| three-bed basket |
| three-bed |
| catalyst bed |
| conversion |
| converter pass |
| ammonia synthesis |
| reaction |
| pressure drop |
| outweight |
| outweigh |
| loop pressure |
| chilling |
| refrigeration |
| refrigeration circuit |
| loop water cooler |
| converter outlet |
| mild |

| converter basket |
|---|
| selected basket |
| selected |
| radial flow |
| converter shell |
| adiabatic |
| interbed |
| inlet gas |
| inlet |
| result in |
| high conversion |
| mechanical design |
| shell |
| piping |
| heat exchanger |

**Conclusion**

Smart dictionaries can greatly improve translators' performance. Developers of such dictionaries should understand that the user deals with the meaning of the texts while the linguistic content is only a means of transferring that meaning. By analyzing the prior history of queries made by all its users, the dictionary can configure its initial filters and prioritization algorithms for further interaction with new users (priors), while the analysis of the ongoing and future sessions allows it to tailor the results to specific users. Future work is aimed at further investigation of these ideas and their practical implementation.

## References

**Kit, Mapκ (2010).** On development of efficient online dictionaries. Based on LexSite dictionary development. *Vestnik RGGU No. 9; Linguistics Series*, Moscow, page 151.

**EUATC (2006)**. The European Translation Markets. *European Union Association of Translation Companies*, Brussels, page 24.

**Kelly, Nataly & Stewart, Robert (2010).** *Common Sense Advisory.*
http://www.commonsenseadvisory.com/Default.aspx?Contenttype=ArticleDet&tabID=64&moduleId=392&Aid=1062&PR=PR.