On the Evolution of Computer Terminology and the SPOT On-Line Dictionary Project

Jiri Hynek, Premek Brada

Department of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Univerzitní 22, 306 14 Pilsen, Czech Republic
e-mail: {jhynek, brada}@kiv.zcu.cz

Abstract

In this paper we discuss the issue of ICT terminology and translations of specific technical terms. We also present SPOT – a new on-line dictionary of computer terminology. SPOT's web platform is adaptable to any language and/or field. We hope that SPOT will become an open platform for discussing controversial computer terms (and their translations into Czech) among professionals. The resulting on-line computer dictionary is freely available to the general public, university teachers, students, editors and professional translators. The dictionary includes some novel features, such as presenting translated terms used in several different contexts – a feature highly appreciated namely by users lacking technical knowledge for deciding which of the dictionary terms being offered should be used.

Keywords: terminology, term, corpus, context, dictionary, language, lexicography, translation, sharing, wiki, English, on-line, information technology, community

1. Introduction

Ordinary users and experts alike get the feeling that Czech translations of computer documentation (including books, help files, reference guides) are of inferior quality [1,2]. It is likely that the same applies to other languages. Readers often prefer original documents in English, which discriminates against those readers that either have no access to these materials, or lack the required linguistic knowledge.

It is generally believed that what makes a translator's work difficult is the technical terminology. But in reality, the main problem is to grasp the *meaning* of specific terms, the actual thing that is denoted by the term, use of the term in the *context*, occurrence of the term with other terms, and the term's stylistic features.

The situation is complicated by the fact that translations are created (literally made up, invented) by technical staff and software developers who lack linguistic skills and „feeling" for the natural language; or at the opposite end of the spectrum, by „professional" translators who lack skills and terminology in the specific ICT (information and communication technology) sub-domain.

Complaints about the poor quality of translations are often backed by the lack of uniformity in addressing specific technical terms. In other words, different authors give the same thing different names. Activities aimed at standardizing computer terminology are very rare; one of the very few is Microsoft Terminology Translations [3] available for 59 different languages at the time of writing. The glossary provided in the form of a CSV file contains more than 12,000 English terms plus the translations of the terms. Unfortunately, translated terms are not provided for all the terms in the list, and some translations seem to be controversial.

In this paper we would like to discuss the issues which are encountered when translating terms arising in a rapidly developing domain like the ICT, including taxonomy of the terms in view of their development status. Secondly we present a project of an on-line dictionary aiming at supporting the work of translators and localization developers. The structure of the paper is as follows. In the following section we discuss the work of translators and the tools they have available. Section 3 describes the taxonomy of terms and some issues with respect to translation. Section 4 presents the SPOT on-line dictionary, the goals of the project, key distinguishing characteristics of the tool, and its current status. The paper is finished with a conclusion.

2. Dictionaries and Resources for Translators

We are living in the age of an information explosion. We would need more than one year to study the amount of information produced worldwide within a single day. It is up to translators to cope with all the changes that take place in the field of their specialty. Searching for technical terminology is one of the most time-consuming tasks of every professional translator. Fortunately, the majority of support data can be found by means of search engines, such as Google. Online dictionaries are also available, although the contents of their free versions are often limited, or they are too general for specific purposes. We are happy to observe the trend of utilizing "collective wisdom" in special Internet projects (Wikipedia being the best example), with users selflessly sharing their expertise with their fellows. Our project is one of those.

The SPOT online dictionary presented here (see http://spot.zcu.cz/) is not the only project of its kind at our University. There is the English-Czech GNU/FDL dictionary available at http://slovnik.zcu.cz/online/ and www.wordbook.cz, which is based on the i-spell database [4] – see Figure 1 below.

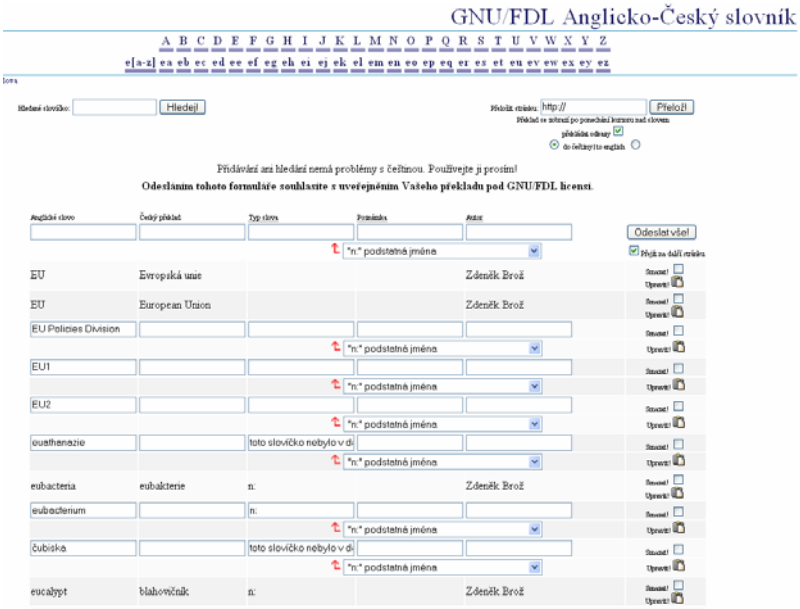

Figure 1: English-Czech dictionary based on the i-spell database

Of course, there are various general dictionaries and encyclopaedias freely available online. These represent an excellent source of linguistic information as well as wisdom, but they are unsuitable for our objective. You can visit the following:

- Merriam-Webster OnLine at: www.webster.com
- Encyclopaedia Britannica online at: www.britannica.com
- Oxford Advanced Learner's Dictionary at: www.oup.com/elt/oald/
- Wikipedia at: http://wikipedia.org/
- Citizendium at: http://en.citizendium.org/
- Dictionary, Thesaurus and Encyclopedia at: www.reference.com
- EuroWordNet (multilingual database with wordnets) at: www.illc.uva.nl/EuroWordNet/
- Free On-Line Dictionary of Computing at: http://foldoc.org/

A comprehensive list of on-line dictionaries for multiple languages suitable for both general and specific purposes can be found at http://a-z-dictionaries.com/online-dictionary.html, or www.yourdictionary.com.

2.1 CAT Tools for Professionals

Today's ICT localization projects often involve millions of words of software documentation, help files, warnings, error messages and other texts to be translated. In order to keep the documentation consistent, memory-based computer-aided translation (CAT) tools are a must. They have a substantial impact on both the translation quality and the productivity of all people involved in the localization project. The roll-out of CAT tools dates back to the beginning of the 1990s. Here are a few examples of today's most popular CAT tools (listed alphabetically):

- Deja Vu (www.atril.com)
- IBM Translation Manager (www.ibm.com)
- MetaTexis (www.metatexis.com)
- SDL Trados (www.trados.com)
- SDLX (www.sdl.com)
- Star Transit (www.star-ag.ch)
- Systran (www.systransoft.com)
- WordFast (www.wordfast.net)

Additional resources for translators and linguists can be accessed via www.multilingual.com, where you can find a collection of more than 1600 links divided into 41 categories (such as Automated Translation, Dictionaries, Internationalization Tools). There are also conferences dedicated to translation and localization industries, such as Localization World (www.localizationworld.com). The localisation community is also supported by The Localisation Research Centre at University College Dublin (www.localisation.ie). Useful links to various translation resources can be found at www.translation.net/links.html. You can download various glossaries at the translators' directory Go Translators (www.gotranslators.com). An example of a free memory translation database for multiple languages can be accessed at www.open-tran.eu.

3. English ICT Terms: From Old-timers to Troublemakers

We are witnessing rapid development in the ICT domain, with quickly changing terminology as one of the consequences. And the language of ICT professionals is suffering badly. The impact on other languages is serious – in the case of Czech, for example, the majority of terms are more or less adopted directly from English. Some terms find their Czech equivalent immediately, some are developing, and for some we still have no equivalent.

For those that enjoy categorization, the following classes of computing terms can be identified.

3.1 Old-timers

This category includes many terms dating back to the 1960s. Their meaning has settled, as well as their translation into other languages. We can subdivide into:

- Morally obsolete terms (such as *bubble memory*, *punch tape*, *punch card*, and recently, also *floppy disk* and *diskette*);
- Stabilized and commonly used terms (such as *display*, *plotter*, *button*, *printer*, *mail server, dialog window*, *RAM, search engine*).

3.2 Novas and Supernovas

English is quite a flexible and even playful language with the ability to create and absorb new words (e.g. *text me* for "send me an SMS message", or abbreviations like *B4*), and new terms therefore appear very quickly, sometimes with an associated hype. Other languages need not be that flexible, which then poses problems in translating these new words, often created ad hoc. Unfortunately, this happens more than often in our mother language.

Compared to the category above, the number of terms labelled as Novas is relatively small. More difficult is the decision taken by a translator or a publication/magazine editor as to how to have these localized. This category includes: *cookie*, *spam*, *phishing*, *blog, freeware*, *emoticon*, *code closure, social bookmarking* and so on.

3.3 Troublemakers

The category of Troublemakers is relatively large. Thanks to Troublemakers, computer dictionaries get published and sold in large volumes, electronic dictionaries flourish on the Internet, and long debates among academics and language purists are held. It is up to computer users, editors and authors which terms shall prevail and which shall become obsolete. We can, again, subdivide into:

- **Homonymous Troublemakers** – identical computer terms with different meanings, such as *collector* (of a transistor) and (software) *collector* (e.g. data collector, portable collector representing a SW component), or *plug-in* (meaning an amplifier) vs. *plug-in module*, a SW component);

- **Synonymous Troublemakers –** The same (or virtually the same) meaning expressed differently, sometimes erroneously (such as *cross reference* vs. *cross-index* vs. *link*, *local menu* vs. *context menu* vs. *pop-up menu* vs. *shortcut menu*, *pull-down menu* vs. *drop-down menu* (or a *list*) *submenu* vs. *child menu, tool palette* vs. *toolbox, custom* vs. *personal, scroll bar* vs. *slide bar, clickable map* vs. *interactive map*).

3.4 Terms on the Move

- Over time, some IT terms **shift and/or extend their meaning**, such as the former *monitor*, meaning today's *screen*, and which now commonly refers to SW utilities; *link,* formerly used in the context of network connections (e.g. *link control protocol*) is now more often used for *web links* or *object linking*;
- **Everyday terms acquire new meaning** in the context of IT: e.g. *signature* is now commonly used in *method signature*, *virus signature*, or *digital signature*; *little endian* and *big endian* (adopted from Gulliver's Travels by Jonathan Swift) now refer to the method of storing multi-byte data; *pool* acquired the meaning of *fund*, i.e. a source of something (*thread pool*, *resource pool*); *key* becomes an "identifier" (*database primary key*); *root* (such as in *tree root*) becomes a type of user, or is used for *root folder*s; *builder* has a new meaning, such as in *application builder*, *list builder*, or *expression builder;* the word *seamless* has become a buzzword in the context of application integration; *heap* currently represents a type of memory; *thread* today refers to a sub-process (such as in *multithreading*); *stamp* has acquired a more abstract sense; *field* (such as in *sports field* or *mine field*) got the meaning of *entry/item*; *host* no longer refers just to persons, but also servers (either hardware or software); *wizard* is commonly used in *installation wizard* or *test wizard*; *garbage* now refers to meaningless data or data no longer needed; *docking* is used for toolbars or laptops instead of (space) ships,
- We are witnessing the process of **heavy verbalization**, e.g. *to right-click* (in place of "click the right mouse button"), *to cache* (in place of "save in cache"), *to host* ("to act as a host"), *personalize*, *televise*, *deserialize,* or *visualize*.

3.5 Esoteric Terms

- **Esoteric IT terms** are used by a relatively narrow group of IT specialists; these include, for example: *setter* and *getter* (in object-oriented programming), *undeploy*, *abstract factory*, *uptime*, *design pattern*, *marshalling*, *serialization*, *proxying*, *entity bean*, *refactoring*, *tight coupling*, *box model*, *locale*, and *hot-swap*.
- **Application- or corporation-specific terms**: these are quite special purpose, found in a limited number of applications, or used only within specific corporations or teams. For example: *rolling period*, *purge*, *context root*, *governance, updater, Apple menu,* and *Start menu*. Please note that these terms may have different meanings depending on the organization/application.

Both situations in which esoteric terms arise, summed by the "narrow user group" characteristic, make their localization complicated, since specialists are happy to use the original (non-translated) terms, while translators lack the technical background knowledge to make up suitable translations.

4. SPOT On-Line Dictionary

Aware of the issues discussed above, the authors started a new project to overcome the difficulties in creating suitable translations in the rapidly changing ICT domain. Our objective is to help the community of translators (plus editors and IT professionals) to either quickly find the correct translation for new, unusual, or tricky terms, or – if no such translation is known – to create one with the assistance of the "collective wisdom" of their colleagues.

Our work is novel in several aspects. Although there are various general-purpose dictionaries available that help in achieving the first objective, we are focusing strictly on the area of ICT, which undergoes frequent changes from the linguistics point of view. The SPOT dictionary is initially starting with a well-established terminology corpus, and its further development will be supported by our extensive experience in localization projects, besides the knowledge of leading IT professionals and editors.

Our approach is also new in that we "settle" the final version of Czech translations within the community of users, under the supervision of ICT specialists. It is anticipated that we will capture the latent interest of the community of translators, readers, editors and companies involved in localization projects (such as translations

of computer publications, documentation and help files, translation of game scripts, and office software localization).

While we believe that professional editorial supervision is absolutely essential for acceptable "settling" of translated terms, the role of the mass of users is seen as a key differentiating point to classical approaches in translation and also as a critical success factor. Similarly to several successful Web 2.0 projects like Wikipedia (www.wikipedia.org), the SPOT project thus hopes to bring into fruitful cooperation professionals and users alike.

Finally, we also offer new features facilitating the work of large or distributed localization teams. Controversial terms are always on-line, rather than stored locally with a localization team member. It is essential that everyone uses the latest version of translated terms. SPOT also eliminates the need to redistribute up-to-date versions of dictionaries, which gets time-consuming in the case of large translation projects.

Last but not least, the added value of our dictionary is greatly enhanced by showing translations in various *contexts* based on on-line search of the Internet. The user can see which of the translations offered should be used, as it can be derived from the context information shown (see Figures 3, 4, 5 below).

4.1 The Dual Role of SPOT

As suggested by the previous paragraphs, SPOT will serve two complementary purposes: a reference dictionary of computer terms, and a platform for "settling" these terms.

4.1.1 Reference dictionary

The basic function of this Internet dictionary is to provide Czech translations for specific ICT terms, either by browsing or by searching (see Figure 2 below). The quality of the translations is guaranteed by the initial English-Czech corpus based on the English-Czech ICT Dictionary written by the first author. Opportunities to provide valuable add-ons on top of the initial corpus are given by the fact that the dictionary is on-line.

Figure 2: SPOT interface

Firstly, editors can assign categorization *tags* to individual terms, in a manner similar to some popular web services such as Flickr for photographs [5] or StumbleUpon for website links [6]. Users will thus be able to confine their search to a specific ICT area, or study other terms related to their area of interest:

- Algorithms and Programming
- Artificial Intelligence
- Communications
- Computer Graphics
- Computer Linguistics, NLP (Natural Language Processing)
- Computer Modelling and Simulations
- Computer Networks and Distributed Systems
- Cryptography, Cryptanalysis
- Data and knowledge mining
- Database Systems
- DTP, Pre-press
- Electrical Engineering and Electronics
- Hardware
- Internet and Web Technologies
- Man-machine Interface
- Mobile Devices
- Operating Systems
- Programming Languages
- Robotics
- Software

Secondly, we can enrich the information about the terms and their translations by displaying their occurrences in various *contexts* based on on-line search of the Internet. We use *Google search API* to obtain on-demand results, configured so that only sites with high relevance to ICT are searched. The probability that the context results will be meaningful is thereby increased. Having spent many years working on large localization projects, we are aware that showing a term's usage in the context is very useful and highly appreciated, especially by translators.

In its current implementation, SPOT can display the following context information (see Figures 3, 4, 5):

- Web sites, no restriction,
- IT webs only,
- Wikis,
- On-line dictionaries,
- Blogs (to be implemented).

A list of authoritative IT web sites, wikipedias, on-line dictionaries and blogs is maintained by the system's administrator.

Disadvantages of the web-based corpus

Corpus gathered from the web is not always reliable and Google cannot provide information that linguists would like to have. In addition,

- Users have no control over the content,
- Web sites are full of metadata that are unnecessary for corpus building,
- We need to focus on pages in a specific language only,
- Specific jargon is used in web blogs and chat rooms.

In spite of the above, we believe that the advantages of providing the user with context information outweigh the drawbacks. Searching for context information via Google is exactly what most translators groping in the dark would do in the first place.

Context information prepared ahead – the Database Corpus

Initially, we were planning to build the corpus manually by collecting context information from the web, selecting the best samples manually and storing the resulting corpus in a text database. We were attracted by the design of the WebBootCaT (see http://corpora.fi.muni.cz/bootcat/ or http://sslmit.unibo.it/~baroni/), which is a tool for building domain-specific corpora to support translators [7].

For more information on web corpora building, see also KWiC Finder (www.kwicfinder.com) ("Web as Corpus"), WebCorp (www.webcorp.org.uk), or Web as Corpus Toolkit (www.drni.de/wac-tk/).

There are two reasons why we rejected the idea of corpus building within a database:

1) SPOT users are on-line and it is unlikely that corpus building on the fly would cause any time delays; context information gathered via Google search API is presented instantly; using a database involves additional overhead plus necessitates database maintenance.
2) ICT terminology evolves very quickly and new terms are created on virtually a daily basis; it would be prohibitive to maintain database corpus up-to-date for this field of specialty.

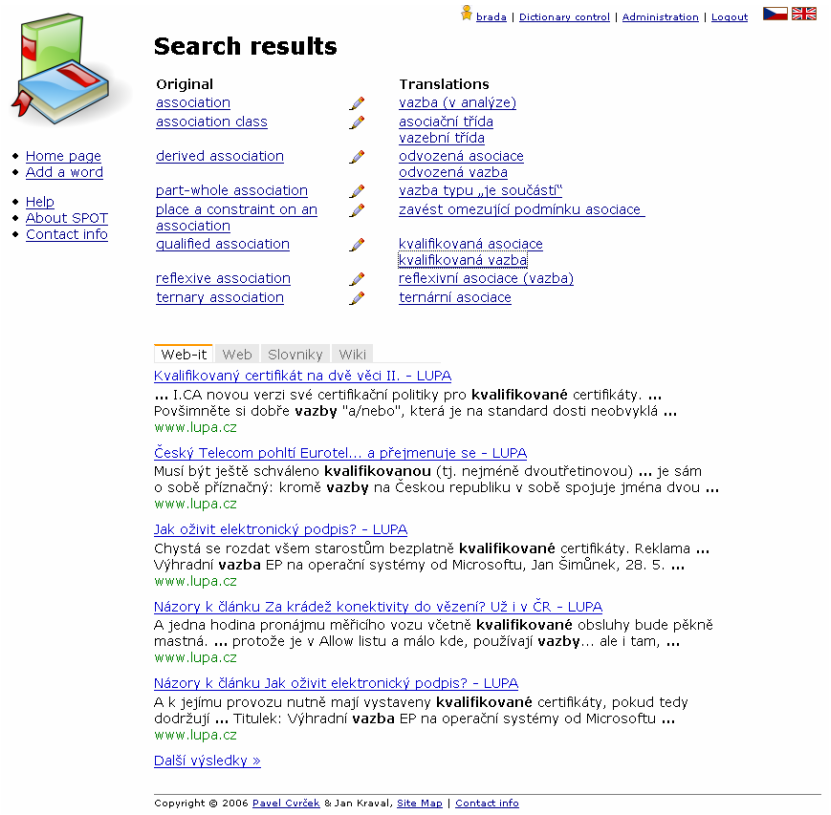

Figure 3: Term translations shown in the context of the web, confined to IT sites only

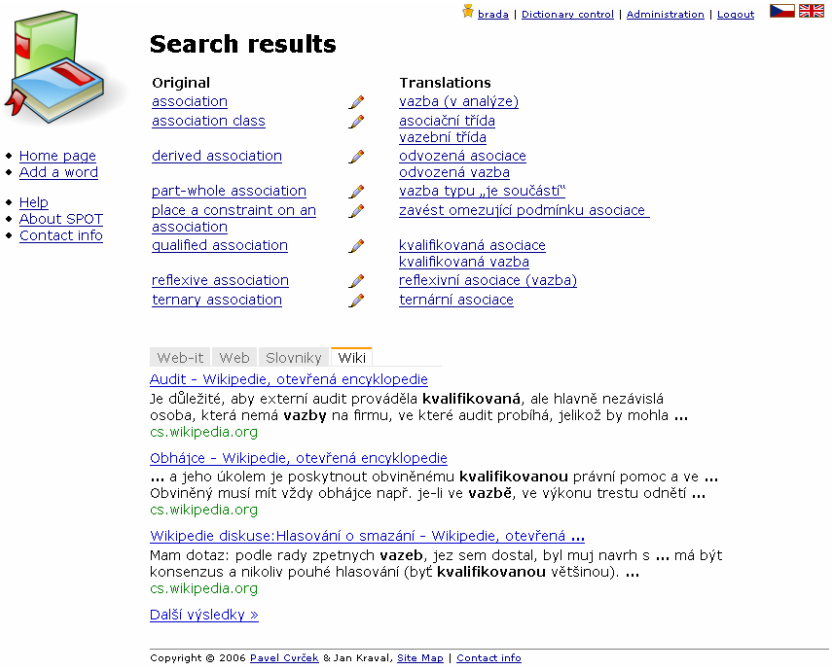

Figure 4: Term translations shown in the context of pre-defined wikis

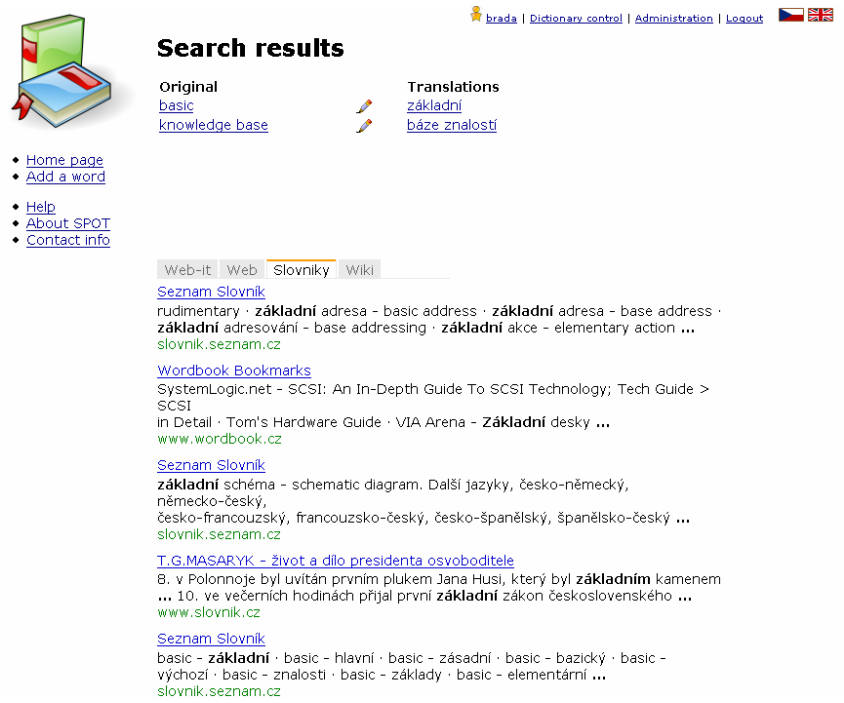

Figure 5: Term translations shown in the context of other on-line dictionaries defined by the system's administrator

Last but not least, there is a set of *minor features* that can be helpful for SPOT users. Apart from the translation(s) deemed correct, the dictionary can show incorrect or unsuitable translations of a term when such were labelled by the editors. This information can be a valuable guide for translators and learners alike, helping them avoid common mistakes. Also, since registered users are able to vote for the translations, popular and well-accepted Czech equivalents become easily visible.

4.1.2 Platform for terminology "settling" by voting

The main advantage of SPOT is that it can act as a platform for the natural development of quality Czech equivalents to the original English terms. With very little effort it can be adapted for any language.

Internet facilitates quick and efficient communication, and it is the inherent property of "collective wisdom" that we are planning to utilize. SPOT will let its users propose Czech translations of unlisted or "unsettled" terms, vote for these translations and discuss them. Based on our long-term observations, users that are likely (and willing) to suggest language equivalents are those that are truly concerned about their form, such as professional translators, professional engineers authoring technical documentation, or academics.

As a necessary complement, the dictionary also includes features for editors to decide upon the final version of individual translations (see Figure 6 below). The final choice on the correct vs. wrong translations will be taken by a small team of editors chosen from the most renowned professionals in the field of ICT, possessing sufficient linguistic knowledge.

Figure 6: Editing, updating and commenting on term translations

The process of establishing an accepted term translation is as follows:

1. Users add (or import from a CSV file) original terms (in English), assigning them to specific categories;
2. Other users propose additional suitable translations, possibly with references to sources of occurrence, and explanatory comments;
3. Registered users can vote on translations and discuss controversial ones;
4. When the discussion has settled, the editor marks the most suitable translation as "Official", and the remaining versions as either "Usable" or "Unusable / Non-recommended".

Since anyone can become a registered user, SPOT supports the idea that translations may become shared work, and consequently a shared responsibility of the "ICT general public". This is a different approach than the prevailing practice, where a few linguists work on the official localization terms in isolation.

SPOT will also find practical use amongst members of localization teams while creating specialized dictionaries for specific translating projects (see Figure 7). The development of a local corpus proceeds as follows:

1. A special section in SPOT is reserved for the localization team (specific project / customer / product);
2. On-line "settling" of controversial translations (editing, voting) takes place within the team;
3. Terms are propagated instantly into the rest of the dictionary corpus; nonetheless, project-specific terms are designated as such. There is a feature to export this partial corpus to other formats, such as CSV, and to import the terms into the team's Computer Aided Translation tool – CAT.

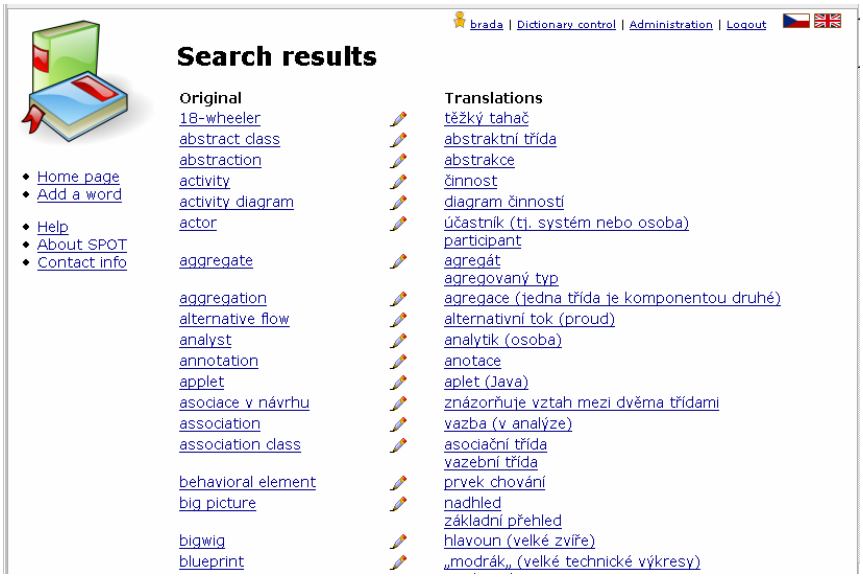

Figure 7: Listing terms filtered by a translation project

The advantages this usage of SPOT brings to the localization teams are manifold, rooted mainly in its on-line implementation:

- Controversial terms are always either on-line, or directly in CAT, rather than stored locally with a localization team member;
- Everyone uses the latest (i.e. currently agreed) version of terms;
- No need to redistribute up-to-date versions of dictionaries.

Below you can see an extract from the MARTIF file used for interchanging newly translated terms among team members (export from Termstar CAT tool while working on a localization project):

```
<langSet lang='eng-us'>
<ntig><termGrp>
  <term>Add a Favorite Place...</term>
  <termNote type='termType'>full form</termNote>
  <termNote type='TS_CreateId'>Jiri Hynek</termNote>
  <date type='origination'>20060917T161007Z</date>
  <termNote type='TS_UpdateId'>Jiri Hynek</termNote>
  <date type='modification'>20060917T161007Z</date>
</termGrp></ntig>

</langSet>

<langSet lang='ces-cz'>
<ntig><termGrp>
  <term>Přidat oblíbené místo</term>
  <termNote type='termType'>full form</termNote>
  <termNote type='TS_CreateId'>Jiri Hynek</termNote>
  <date type='origination'>20060917T161007Z</date>
  <termNote type='TS_UpdateId'>Jiri Hynek</termNote>
  <date type='modification'>20060917T161007Z</date>
</termGrp></ntig>
```

4.2 SPOT Implementation – The Current Status

We have currently implemented basic features that include searching, adding new translations, tagging and commenting, dictionary administration by the editorial team, and namely showing translated terms in the context. Under way is the implementation of discussion forums to translations, as well as voting for individual terms suggested by users or editors. Cross-referencing (“see also” links) will be featured in a subsequent release as well. The application is implemented on the Java 5 platform, utilizing the Spring framework and PostgreSQL database system.

The SPOT dictionary can be accessed at http://spot.zcu.cz, where interested readers will also find links to project-related pages. For the data model of the current implementation, see Figure 8 below.

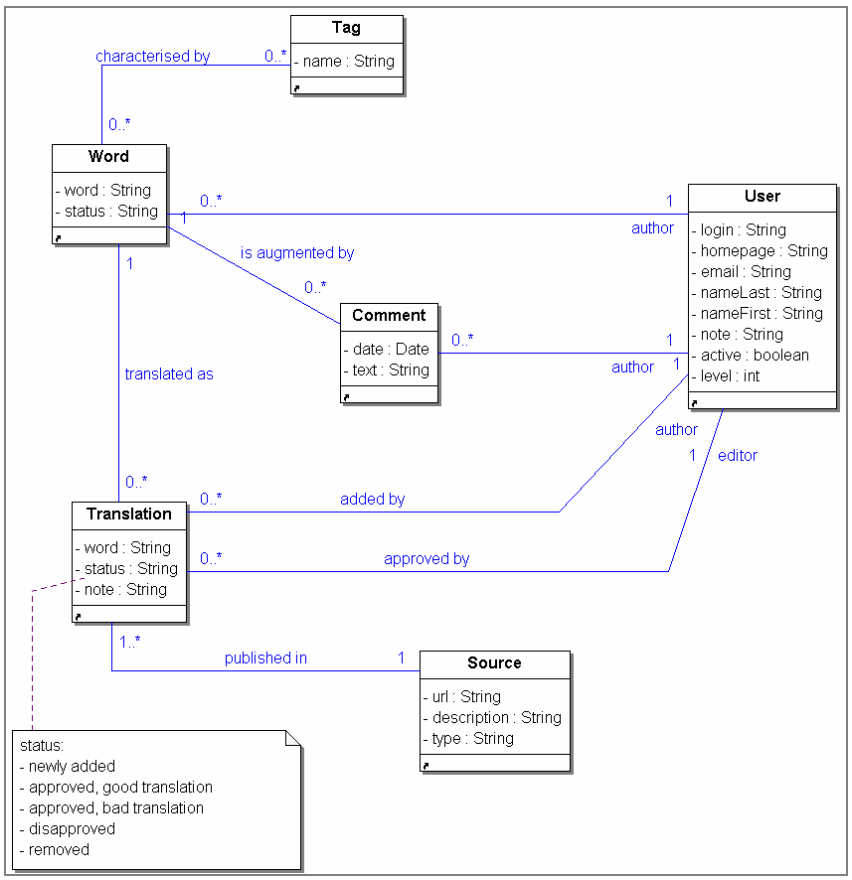

Figure 8: SPOT data model

5. Conclusion

It is not the task of university teachers or pure linguists to propose suitable equivalents of specialized English terms in their native language. Rather than these people, it should be up to the editors of computer magazines, students, and translators of IT documentation. Terminology must adapt to users, not the other way round.

We believe that by implementing the solution proposed herein, we can not only improve the quality of computer documentation translated into Czech, but also the general culture in this area. Indirectly, this may help to increase readers' preferences for translated publications, while curtailing discrimination against those who either do not have the means to obtain the original books, or the language skills to read them.

Hopefully, SPOT will help in at least the partial standardization of terminology being used, and become a useful source of information for persons involved in the technical or academic writing process.

6. Acknowledgements

This work was partly supported by the Ministry of Education of the Czech Republic under Grant No. 2C06009 within the National Program for Research II.

7. Notes and References

[1] Virius M. On the quality of Czech translations of technical literature (in Czech: Nad kvalitou českých překladů odborné literatury). Proceedings of OBJEKTY 2002. Praha 2002. ISBN 80-213-0947-4.
[2] Mondschein P. Oficially dishonoured Czech (in Czech: Oficiálně prznėná čeština.) http://games.tiscali.cz/specials/prznenacestina/index.asp. Published 8th September 2004.
[3] Microsoft Terminology Translations, available at: www.microsoft.com/globaldev/tools/MILSGlossary.mspx

[4] The i-spell database, available at: www.lasr.cs.ucla.edu/geoff/tars/
[5] Flickr – Online photo management and sharing application, available at: www.flickr.com/
[6] Stumbleupon – Personalized recommendations of websites, videos, pictures and more. Available at: www.stumbleupon.com/
[7] Baroni, Kilgarriff, Pomikálek, Rychlý. 2006. WebBootCaT: instant domain-specific corpora to support human translators. In: Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation. Oslo, Oslo University (Norway), ISBN 82-7368-294-3.