GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Bratislava: SNK JÚĽŠ 2004. Dostupné z: <http://korpus.juls.savba.sk/publications/block2/tokenizacia-lematizacia-a-morfologicka-anotacia-slovenskeho-narodneho-korpusu/Tagset-aktualny.pdf.>

GIANITSOVÁ, Lucia: Zamyslenie nad výučbou zámen a čísloviek pri príprave morfologickej anotácie SNK. In: Tradiční a netradiční metody a formy práce ve výuce českého jazyka na základní škole. Sborník prací z mezinárodní konference konané 19. 3. 2004 na Pedagogické fakultě UP v Olomouci. Ed. M. Polák – K. Vodrážková. Olomouc: Univerzita Palackého 2005, s. 53 – 65.

Krátky slovník slovenského jazyka. 4. vyd. Red. J. Kačala – M. Pisárčiková – M. Považaj. Bratislava: Veda 2003.

Morfológia slovenského jazyka. Red. J. Ružička. Bratislava: SAV 1966. 896 s.

Parallel Corpus of Computer Terms

Radovan Garabík

Ľudovít Štúr Institute of Linguistics Slovak Academy of Sciences, Bratislava

1. Introduction

Corpora play an important rôle in modern linguistics[1], a situation greatly facilitated by current boom of cheap and widely available computing power. Parallel corpora form a smaller, but nevertheless important part of corpus linguistics, and have direct utilisation for end users dealing with bi- or multilingual texts.

When compared with „traditional“ monolingual corpora, parallel corpora have several distinguishing features and their creators have to deal with specific problems. First of all, parallel corpora need parallel texts in several languages, which can be sometimes a big obstacle. To get the rights to use texts is inheritably much more difficult than is the case of monolingual corpora[1] , because we need to consider different copyright law(s) in different countries, which by itself is rather difficult subject.

Then there is the question of aligning, using manual alignment is often impractical (even if typical sizes of parallel corpora are of an order of magnitude smaller than typical sizes of monolingual corpora), and writing tools for automatic alignment is far from trivial. On the other hand, parallel corpora are not really expected to have such a detailed and elaborate linguistic markup as monolingual ones, since their main usage and area of interest is shifted away from intrinsic linguistic properties of given language, towards relation between the languages.

[1] while, according to usual copyright laws, it is possible to use texts for educational purposes, it is not really clear if it is possible to make such a corpus publicly accessible

2. Translations in Software Products

Rather recent phenomenon in software world is the existence of *internationalization*[2] and *localization*[3] , which reflect the penetration of computers into many regions of human society, and the subsequent need to use the software either: 1. to work with language (text documents, DTP, databases) other than English, and 2. by people without adequate command of English. Point (1) is the aim of internationalization – modifying software to be able to deal with languages other than English, point (2) of localization – making software communicate with users in their respective languages. Of these, we are particularly interested in localization, because it implies the necessity of translating user interface(s) into targeted languages.

Unfortunately, in case of commercial software, to get the texts in electronic form suitable for inclusion into a corpus is probably even more difficult than with monolingual corpora. However, there exists a lot of translated software under different OpenSource licenses, such as GNU General Public License, GNU Lesser General Public License, BSD License, Artistic license, X11 license and derivates, and these kinds of licenses allow us to include the translations without hindrances.

The standard system and API for translations in OpenSource world is the GNU gettext[2] system, although there are numerous exceptions. Using gettext computer translations has many advantages – the most important for a parallel corpora is the fact that the translations are perfectly aligned on expression level.

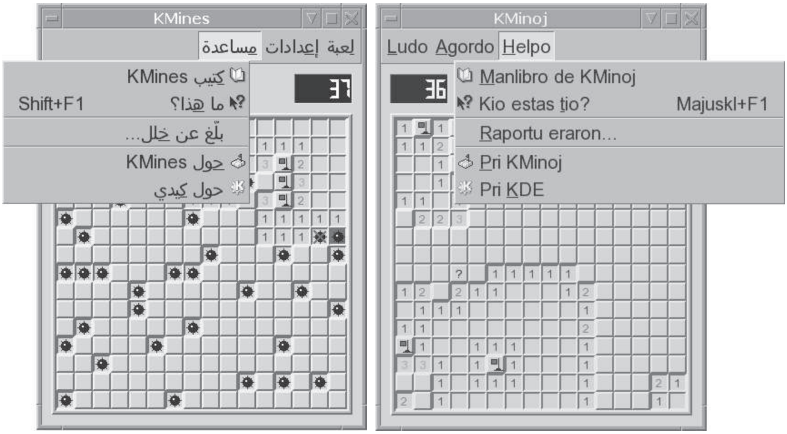


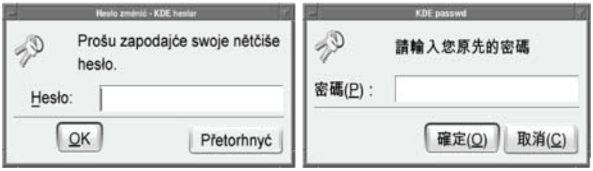
Figure 1: Example of a popular software translated into different languages. Notice how the translation affects the direction of menus and icons.

[2] often shortened to *i18n*

[3] or *l10n*

Using such a specialised area of translations, we have to be aware of several consequences. Overwhelming majority of translations (as of the software in general) is prepared by amateurs in linguistic profession, though professionals, or at least highly skilled in software engineering. This is strikingly different from translations of commercial software, where most translations are done by professionals in linguistic area, but untrained in computer skills, and therefore often unaware of true meaning of texts being translated (but, to be fair, often with the help of consulting computer specialists).

This can be seen as both the advantage and a disadvantage. Disadvantage, because the quality of translations is often very poor, with many mistakes and mistranslations. Advantage, because it better reflects the actual use of language by computer specialists, not as prescribed by institutional bodies and norms, often disconnected from real life[4] .



Hesło změnić - KDE hesłar

Prošu zapodajće swoje nětčiše hesło.

KDE passwd

請輸入您原先的密碼

Figure 2: Password change program in two different translations.

3. Sources of Translations

We used KDE[3], GNOME[4] and GNU Translation Project[5]. There are often several versions of a given software, sometimes with changes in UI, which are further reflected in translations. We decided to include the translations from the older versions as well, because it provides more translated expressions. The only slight disadvantage is the fact that new versions are often an impulse for translators to correct mistakes in their previous translations, and by keeping the old ones we keep also the less correct ones.

Using the above mentioned sources also means that the original language is always English. In fact, software with first language other than English is very difficult to find.

4. Connecting Translations

As an example, let's take original term „File“. This is present in almost every GUI software, but can have different translations, according to contexts. For

[4] For example, in the Slovak expressions of current version of the parallel corpus, there is exactly one (1) occurrence of (otherwise officially prescribed) Slovak word „lomka“, whereas the (officially forbidden) alternative „lomítko“ occurs 28 times.

example, it has been translated (in different programs) into one target language as „Súbor", „SÚBOR", „file" and „zadaného súboru", respectively, while into the second target language it has been translated as „Fajl", „ДАТОТЕКА", „Датотека" and „Фајл". We decided to retrieve expressions up to second level of connections, e.g. for user query „zadaného" we get matching expression „zadaného súboru", which in turn corresponds to original expression „File", giving back to user the results „Súbor", „SÚBOR", „Fajl" ... The situation is illustrated on picture 3.

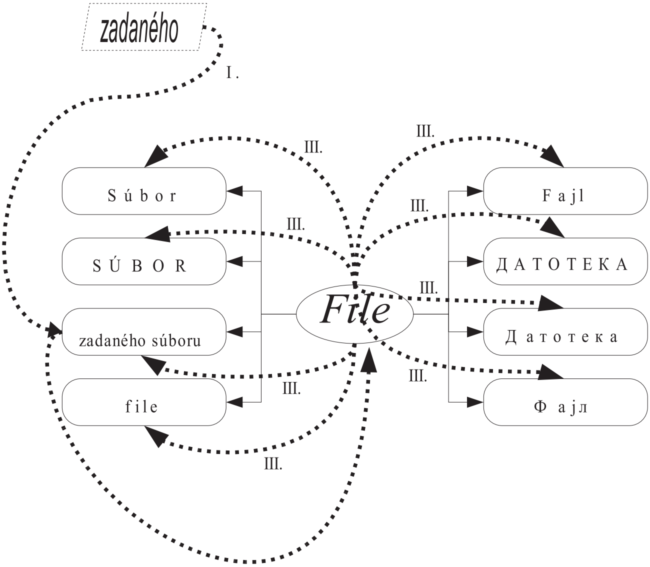


Figure 3: Full lines show the connections between original and translations, dotted lines the path of user query to a corresponding expression (I.), then to the original expression (II.) and back to all available translations (III.)

5. Plurals

Good internationalization of plural handling is surprisingly difficult from computers' (and programmers') point of view. We do not need to go into lengthy descriptions about languages having only singular, having dual, paucal and similar grammar categories – this has been described elsewhere[6, 7]. We just need to realize that the decisive factor for a computer system is the specific *form* of textual unit, not the division into grammar categories[5]. Just by looking at Slavic languages, we see

[5] This is trivial and obvious for computer scientists and programmers, however this notion is rather unnatural for linguistically oriented persons. We felt obliged to elaborate on this.

different noun form for numeral 1, then a different form for 2, 3 and 4, a different form for numerals greater or equal than 5, and often repeating scheme for numerals greater than 20, according to modulo 10, with many exceptions.

While and English-speaking program uses simple piece of code like this,

```
print „Uploaded", number,
if number==1:
  print „file"
else:
  print „files"
```

typical Slavic-speaking equivalent would be lengthy and complicated. On the other hand, similar Chinese- or Hungarian-speaking program would be rather simpler. These issues have even more serious effects on internationalization, since we cannot reasonably expect program authors to know and use all the different rules for many different languages, not speaking about bloat and unmaintainability of such software.

GNU gettext solves the situation by requiring explicit enumeration of different forms, and by providing a rule for choosing the appropriate form. So the expression would be

```
Plural-Forms: nplurals=1; plural=0;
```

for Hungarian,

```
Plural-Forms: nplurals=2; plural=n != 1;
```

for English, and

```
Plural-Forms: nplurals=3; \
plural=(n==1) ? 1 : (n>=2 && n<=4) ? 2 : 0;
```

for Slovak.

Situation gets even more complicated as wee move into more exotic languages. We decided to keep things simple and we are grouping all the possible plural forms together, so for example query „súbory“ would match „súbor“ and vice versa.

6. Technical Details

Each expression has a unique ID number, and all the expressions are stored in a MySQL database. Two tables are used, the first one keeps the connection between expression and its ID, information about whether the expression is an original or a translation, and the language of given expression.

The second table contains pairs of expression number (be it original or a translation) and a link to corresponding translated (or original) expression. The lookup consists first of finding list of IDs matching user query (with optional restriction about expression language), and then for each of these ID a list of corresponding linked expressions is retrieved and transformed into expressions again. The search system has simple WWW cgi-bin inteface (fig. 4) providing a way to search for a substring in the expressions, with a possibility to limit the search only for „interesting“ languages. Everything is implemented in the Python programming language.

Though MySQL is probably the worst solution for corpus backend storage[6], and we are leaving the searching for the substring to internal MySQL processing, which add up tot the overall inefficiency, the combined power and ease of use of Python and MySQL lead to very quick deployment of the whole corpus. As an additional feature, it is possible to update the expressions on the fly without taking the corpus down or reindexing it. The speed of resolving queries is quite acceptable, but the improvements (especially by using more corpus manager-like backend) are planned.

The corpus is publicly available via Slovak National Corpus WWW page[7].

7. Statistics

Currently, the corpus contains 1.6 million different expressions in 88 languages. The number of words is more than 11.5 million – many scripts do not use any separators between words, we have counted only those that are separated by whitespace and common interpunction, so the number will be probably noticably higher. Only counting characters in Chinese expressions, we get additional million of „words“, if we can consider one chinese character to be one word. There are additional 2.3 million characters in Japanese expressions, however one Japanese word typically consists of several characters. And we are not speaking about other languages with scripts without word delimiters – though Japanese and Chinese are probably the most prominent ones.

Following table shows for each langugage the number of expressions present in the corpus, relative count with respect to the whole number of expressions, and relative count with respect to the number of English expressions (in other words, how many % of original expressions have been translated).

ISO 639	Language	№ expressions	rel. [%]	rel.to English [%]
af	Afrikaans	5094	0.30	3.30
am	Amharic	4306	0.26	2.79

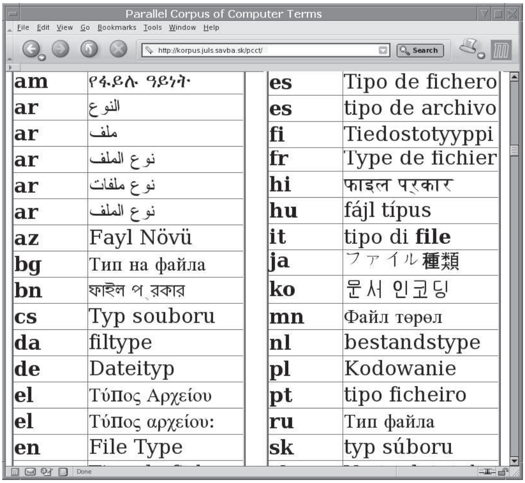
[6] in author’s opinion, the second worst is the XML format

[7] http://korpus.juls.savba.sk/

ISO 639	Language	№ expressions	rel. [%]	rel.to English [%]
ar	Arabic	52109	3.10	33.74
az	Azerbaijani	18868	1.12	12.22
be	Byelorussian	19098	1.14	12.37
bg	Bulgarian	10732	0.64	6.95
bn	Bengali; Bangla	11663	0.69	7.55
br	Breton	289	0.02	0.19
bs	Bosnian	5027	0.30	3.26
ca	Catalan	41313	2.46	26.75
cs	Czech	32984	1.96	21.36
cy	Welsh	15448	0.92	10.00
da	Danish	59991	3.57	38.85
de	German	63785	3.79	41.30
el	Greek	25452	1.51	16.48
en	English	154430	9.19	100.00
en_GB	English (Great Britain)	17957	1.07	11.63
eo	Esperanto	8364	0.50	5.42
es	Spanish	69736	4.15	45.16
et	Estonian	38381	2.28	24.85
eu	Basque	8193	0.49	5.31
fa	Persian	6581	0.39	4.26
fi	Finnish	27884	1.66	18.06
fo	Faeroese	1702	0.10	1.10
fr	French	74611	4.44	48.31
ga	Irish	6995	0.42	4.53
gl	Galician	12036	0.72	7.79
he	Hebrew	18293	1.09	11.85
hi	Hindi	11197	0.67	7.25
hr	Croatian	7902	0.47	5.12
hu	Hungarian	31723	1.89	20.54
ia	Interlingua	70	0.00	0.05
id	Indonesian	12162	0.72	7.88
is	Icelandic	5964	0.35	3.86
it	Italian	40189	2.39	26.02
ja	Japanese	39188	2.33	25.38
kn	Kannada	1380	0.08	0.89
ko	Korean	21143	1.26	13.69
ku	Kurdish	1198	0.07	0.78

ISO 639	Language	№ expressions	rel. [%]	rel.to English [%]
lg	Ganda	202	0.01	0.13
li	Liii	5067	0.30	3.28
lo	Laothian	2658	0.16	1.72
lt	Lithuanian	19014	1.13	12.31
lv	Latvian, Lettish	15095	0.90	9.77
mi	Maori	310	0.02	0.20
mk	Macedonian	17600	1.05	11.40
ml	Malayalam	4887	0.29	3.16
mn	Mongolian	17972	1.07	11.64
mr	Marathi	687	0.04	0.44
ms	Malay	15499	0.92	10.04
mt	Maltese	5069	0.30	3.28
nb	Norwegian (Bokmål)	4854	0.29	3.14
ne	Nepali	2435	0.14	1.58
nl	Dutch	38978	2.32	25.24
nn	Norwegian (Nynorsk)	15263	0.91	9.88
no	Norwegian	16176	0.96	10.47
nso	Northern Sohto	3685	0.22	2.39
oc	Occitan	678	0.04	0.44
pl	Polish	32856	1.95	21.28
pt	Portuguese	33194	1.97	21.49
pt_BR	Portuguese (Brasil)	38004	2.26	24.61
ro	Romanian	24548	1.46	15.90
ru	Russian	43281	2.57	28.03
se	Northern Sámi	4823	0.29	3.12
sk	Slovak	43869	2.61	28.41
sl	Slovenian	28411	1.69	18.40
sq	Albanian	13000	0.77	8.42
sr	Serbian	19112	1.14	12.38
sr@Latn	Serbian (Latin)	14982	0.89	9.70
ss	Siswati	5017	0.30	3.25
sv	Swedish	39819	2.37	25.78
ta	Tamil	9829	0.58	6.36
th	Thai	5593	0.33	3.62
tr	Turkish	40821	2.43	26.43
uk	Ukrainian	22201	1.32	14.38
ur	Urdu	100	0.01	0.06

ISO 639	Language	№ expressions	rel. [%]	rel.to English [%]
ven	Venda	3151	0.19	2.04
vi	Vietnamese	14088	0.84	9.12
wa	Walloon	7969	0.47	5.16
xh	Xhosa	6492	0.39	4.20
yi	Yiddish	1739	0.10	1.13
zh	Chinese	96	0.01	0.06
zh_CN	Chinese (PRC)	28397	1.69	18.39
zh_TW	Chinese (Taiwan)	26368	1.57	17.07
zu	Zulu	3812	0.23	2.47
Total		1681139	100.00	1088.61


Parallel Corpus of Computer Terms
File Edit View Go Bookmarks Tools Window Help
http://korpus.juls.savba.sk/pcct/
Search
Done

am	የፋይሉ ዓይነት	es	Tipo de fichero
ar	النوع	es	tipo de archivo
ar	ملف	fi	Tiedostotyyppi
ar	نوع الملف	fr	Type de fichier
ar	نوع ملفات	hi	फाइल प्रकार
ar	نوع الملف	hu	fájl típus
az	Fayl Növü	it	tipo di **file**
bg	Тип на файла	ja	ファイル種類
bn	ফাইল প্ রকার	ko	문서 인코딩
cs	Typ souboru	mn	Файл төрөл
da	filtype	nl	bestandstype
de	Dateityp	pl	Kodowanie
el	ΤύΠος Αρχείου	pt	tipo ficheiro
el	ΤύΠος αρχείου:	ru	Тип файла
en	File Type	sk	typ súboru

Figure 4: Screenshot showing the WWW interface. Notice several alternative translations existing in some languages.

References

ŠIMKOVÁ, Mária: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Budmerice 2003.

GNU gettext manual. <http://www.gnu.org/software/gettext/gettext.html>

DIEHL, Thomas: The KDE Translation HOWTO. <http://i18n.kde.org/translation--howto/>

The GNOME Translation Project. <http://developer.gnome.org/projects/gtp/

GNU Translation Project. <http://www2.iro.umontreal.ca/~gnutra/po/HTML/>

ČERMÁK, František: Jazyk a jazykověda. Praha: Karolinum 2001.

„Grammatical number“. Wikipedia: The Free Encyclopedia, Wikimedia Foundation Inc. Updated 14 April 2004. <http://en.wikipedia.org/wiki/Grammatical_number>