DIFFICULTIES OF CREATING A LEGAL TERM BASE

Imre Attila, Assoc. Prof., PhD, Sapientia University of Tîrgu Mureş

Abstract: The present paper tries to present the difficulties of creating a Romanian–English legal term base with the aim of integrating it into computer-assisted translation tools. The presented data are taken from seven Romanian–English law dictionaries, taking into consideration both quantitative and qualitative aspects. The quantitative part shows us the number of entries for a specific term, whereas the qualitative part derives from the relevance of the legal terms found. We will discuss specific problems for Romanian (diacritical signs, alphabetical order) and English (typographical errors, spelling issues) or the conversion of the .xls(x) format in a standard one for term bases (.csv).

Keywords: terminology, term bases, legal, quality, Romanian, English.

1. Introduction

Studies regarding terminology prove the importance of updating term bases as new words and expressions continuously enter languages. Living in the age of globalization, localization and the (r)evolution of technology (Imre, 2013) it is no wonder that English as a global language pervades our entire life. In Barber's words:

> "… the onrush of economic and ecological forces that demand integration and uniformity and that mesmerize the world with fast music, fast computers, and fast food – with MTV, Macintosh, and McDonald's, pressing nations into one commercially homogeneous global network: one McWorld tied together by technology, ecology, communications and commerce. The planet is falling precipitantly apart *and* coming reluctantly together at the very same moment." (Barber, 1992, p. 53)

'Mesmerized' by this McWorld, it is easy to observe that it brought about changes in language as well. For instance, Snell-Hornby describes it as "McLanguage", which "is to a great extent a particular brand of American English, reduced in stylistic range and subject matter, and – with the aid of abbreviations, icons, acronyms and graphic design – it is tailor-made for fast consumption" (Snell-Hornby, 2006, p. 132). She continues with the idea of "Eurospeak" of a multilingual continent, where we can talk about a real 'Empire of English' "with the ever-expanding language industry, and here technological aids will continue to play a central part" (Snell-Hornby, 2006, p. 144). This McLanguage of fast consumption in the majority of cases does not meet the requirements of linguists (poor grammar, sub-standard choice of words, etc.), but it may be a real challenge to translators and interpreters. Naturally, there are fields where McLanguage may have little to say (specific languages of medicine, law, etc.), but language mediators should be always prepared to select the best option regarding language register and style.

We should also take into consideration Gouadec's statement from 2007: "The PRAT or Pencil and Rubber-Assisted Translator is clearly on the way out, though there are still a few specimens at large. The Computer-Assisted Translator has taken over." (p. 109). The combination of ideas from Barber, Snell-Hornby and Gouadec already offer the profile of

modern translators, who – in our belief – should have multiple skills and competences: alongside with the two major classical competences (linguistic and cultural), they should be able to deal with the technical revolution (technical competence) and to survive on a global market (self-management).

The present article focuses on the technical competence, more precisely on the importance of translation software, including machine translation, computer-assisted translation (CAT) tools and the ability to handle (open, read, create, modify, encode) various types of text-based documents. Due to the limits of the present paper, we only highlight the importance of term bases (TB), which is a major constituent of all CAT-tools (*SDL Trados Studio, Déjà Vu, Wordfast, OmegaT, memoQ,* etc.).

We have already seen that languages are under a constant change, so monolingual, bilingual or multilingual dictionaries and term bases (containing specific words) need a regular 'update'. Creating and/or finding these databases is crucial during the work of the modern translator, who have to filter the information very effectively in order to select the best option available within a language.

However, in the age of McLanguage quality is a delicate matter (the 'copy-paste' option without paying attention to details is effortless), and for the large public intelligibility settles all disputes. Thus the Romanian expression *pentru că așa vrea mușchii mei* (**because my muscles wants it*) is acceptable, even if ungrammatical.

2. The actuality of a Romanian–English and English–Romanian legal dictionary

We already argued in favor of checking previous dictionaries and term bases as languages are in constant change. This means that even within a specific field new words enter, old ones disappear, or new synonyms and antonyms may appear any time. Furthermore, when more than one language is involved the tendency is much more powerful, especially when one of the languages is English. During the past fifteen years at least ten Romanian – English and/or English – Romanian dictionaries were printed on legal terms, but further dictionaries (especially on economics) also contain valuable entries belonging to legal terminology (e.g. Ionescu-Cruțan, 2006, Năstăsescu, 2006 or Bantaș & Năstăsescu, 2000). Some of them are only Romanian–English (Lozinschi, 2008), while others contain terms in both directions. Thus we can say that there are more dictionaries to choose in the combination of Romanian–English language pair regarding legal terms, although these dictionaries may present large differences regarding the quantity and quality of the terms they contain.

The actuality of a possible term base derives from the above-mentioned globalization and localization, as well as the McLanguage and 'Eurospeak'. Romania is a member state in the European Union, and Romanian is one of the 24 official and working languages. This means that specific languages belonging to a variety of fields (politics, economics, law, management, etc.) are of high interest, and a proper translation memory (or term base) offered by the Directorate-General for Translation (DGT) is of high interest[1]. However, this only contains items that were used in the extensive work of translation into the 24 languages, and it is not within the reach of all translators.

[1] http://ec.europa.eu/dgs/translation/translating/index_en.htm, 29.11.2014.

A much simpler option would be an extensive term base, unifying more available dictionaries in an available format, such as Portable Document Format, but if we have in mind further completion (new entries to be added later), we can create either a Microsoft Office *Excel* file in .xls or .xlsx format or an OpenOffice or LibreOffice *Calc* file in .ods format. In the next section we will present our partial results based on six Romanian–English dictionaries.

3. Difficulties of creating a term base

At first approach to creating a term base the task seems simple: let us take as many dictionaries and glossaries available and transform them into a single electronic file. However, if we have in mind quality above all, we should analyse both the source and the outcome attentively.

3.1. The importance of the source

The importance of the source should never be overlooked, as our partial results presented below signal extremely many types of possible errors. We can even predict that the more languages are involved, the more possibilities to commit mistakes occur. In our research – up to now – we have collected data from seven Romanian–English law dictionaries, having in mind only the entries under letter *Î* (letter *I* with circumflex, one of the five specific Romanian letters).

3.1.1. A proper selection of entries

Specialized dictionaries should contain concepts in specific fields (terms) and they are supposed to be onomasiological (first identifying concepts then establishing the terms used to designate them, cf. Landau, 2001, pp. 1–6), while general dictionaries should be besemasiological (word to definition), although dictionaries cannot really respect this, especially when bilingual dictionaries are involved. A further problem is the definition of *term* itself: "a. A word or phrase used in a definite or precise sense in some particular subject or discipline; a technical expression. B. Any word or group of words expressing a notion or conception, or used in a particular context; an expression (for something. Trumble & Stevenson, 2002). This means that any number of 'words' may form an expression, thus we can even get to the limit of a full sentence within a headword. And this is exactly what happened in our sources:

(Lozinschi, 2008, pp. 263–304)	
***începere**a unui rechizitoriu înainte de a fi precedat de datele *necasare este o încălcare a Convenției de la Viena, art 36* (*instead of *necesare*)	*commencement of an interrogation before the information is given is a breach of the Vienna Convention, Art 36*
*L-au **înhăța**t pe hoț și l-au dat pe mâna poliției.*	*They captured the thief and handed him over to the police.*

Table 1. Sentences in dictionaries

Furthermore, even the combination of words into specialized terms may raise questions, as some readers may think they purely and simply do not belong to a law dictionary:

(Lozinschi, 2008, pp. 263–304)	
Întorsura Buzăului	*the curve of the Buzău* (geographical area in Romania)
(Dumitrescu, 2009, pp. 120–128)	
*înregistrarea titlurilor de proprietate asupra pământurilor (*australia)* (* It should have been capitalized.)	**terrence registration* (not to be found elsewhere, is it so important to be included?)

Table 2. More words as a single term

However, their significance is minor compared to the number of occurrences of this type. A much more serious problem is the mixture of specialized terms with general ones, to be found in all dictionaries. If numbers, countries, body parts are acceptable in legal dictionaries, then all the words existing in a language should be included, as all of them may appear in a particular legal context. This means that lexicology and terminology is not clearly separated (cf. prepositions as headwords), thus any concept, word, term and expression may enter the dictionaries.

The number of irrelevant entries from the point of view of a legal dictionary (*în orice situație* 'whatever the situation'; *în față* 'in front of') thus contributes enormously to the quality of a specialized dictionary as well as its ampleness.

3.1.2. Typographical errors

Typographical errors are probably the most stinging ones in any dictionary("Typographical error," 2014). In an age when human work is aided by computers and software, typographical errors should not be a problem. We may rely on *Microsoft Office Word*'s *Spelling and Grammar* function within *Review* or *LibreOffice Writer*'s *Spelling and Grammar* function within *Tools* (both having F7 as a shortcut), as they can detect the great majority of typographical mistakes[2] (some of them corrected automatically, while typing). However, two conditions must be fulfilled: the proper spellcheck for the particular language must be installed/activated language and the typing language must be set properly. The cases below prove that at least one of these two conditions were not fulfilled while the dictionaries were being edited:

- Extra letter: ***neighbourghood*, ***prolongued*, ***terrified* (Lozinschi, 2008); ***tresspass* (Dumitrescu, 2009);
- Missing double letter: ***accomodate*, ***commital* (Botezat, 2011 and Hanga & Calciu, 2009); ***plaintif* (Voroniuc, 1999 and Voroniuc, 2011); ***înegurat* instead of *înnegurat* (Lozinschi, 2008);
- Missing letter: ***împutenicit* instead of *împuternicit* (Voroniuc, 1999); ***repeted* instead of *repeated* (Lozinschi, 2008);

[2] Except for the 'atomic typos': http://en.wikipedia.org/wiki/Typographical_error#Atomic_typos, 05. 10. 2014.

- Mixed letters: ***ent*re*prise* (Voroniuc, 1999 and Voroniuc, 2011); ***no*si*e* instead of *noise* (Botezat, 2011); ****beseige* instead of *besiege* (Lozinschi, 2008);
- Mistyped letter: ****a*bstinacy* instead of *obstinacy*, ****en*bankment* instead of *embankment* (Lozinschi, 2008); ****set up o company* (Voroniuc, 1999 and Voroniuc, 2011);
- Words stuck together: ****indubio** pro reo* instead of ***in dubio** pro reo* (Hanga & Calciu, 2009);
- Fat-finger syndrome: ****warra*m*ty* (M-N), ****infringem*n*ent* (M-N and extra letter), ****provi*k*e* (I-O), ****ascultare*r*a* (E-R and extra letter) (Lozinschi, 2008);
- Small letters instead of capital ones: ****a*ustralia*, ****u*nited *s*tates *p*atent *o*ffice* (correctly: *United States Patent and Trademark Office*) (Dumitrescu, 2009);
- Capital letters instead of small ones: the title of the dictionary is *Dicționar **R**omân–**E**nglez* instead of *Dicționar român–englez* (although the CIP description is already correct, (Lozinschi, 2008);
- Atomic typos (meaningful words in the "wrong" place, thus spellcheckers will not detect them as errors (Bloom, 2012): *goal* instead of *gaol* ('jail') (Voroniuc, 2011).

In our view, the most grievous types of errors in a dictionary are the ones –whatever type – to be found in the headwords: ****înegurat* (înnegurat), ****încăirare* (încăierare), ****înșeală* (înșală) (Lozinschi, 2008) or incorrect alphabetical order. The latter is extremely problematic in Dumitrescu's book: apparently, entries containing Romanian diacritical marks (*ă, ș, ț* come before any other letter. Thus *înștiințare, înțelegere* comes before *înainta*; *înșelăciune* appears before *însoțitor*; *înstrăinări* precedes *înstrăina*, or we can find *întreținut* in front of *întreaga* (Dumitrescu, 2009, pp. 122–127). In Lister and Veth's dictionary (Lister & Veth, 2010) we could find an entry beginning with *I* among terms beginning with *Î* (*inginerie, ~ genetică*). These errors may turn any dictionary inefficient as the readers may not find the term they are searching for.

3.1.3. Grammatical errors

Grammatical errors mainly derive from ignorance. In case of verbs, we found cases when the conjugated form was used instead of the infinitive: *închirie**z*** ('I rent', first person singular, present) instead of *închiria* ('to lease', 'to rent'). What is worse, there is a separate entry for *închiria*, so this must have been *închirie**re***, which is a noun ('letting', 'renting'). Further errors include singular/plural shift, which remains unmarked: *încasare* ('collection', 'cashing'), whereas an expression rooting from this entry and requiring plural is marked this way: *-i și cheltuieli*, resulting in ****încasarei* instead of *încasări* (Hanga & Calciu, 2009); in other cases instead of a singular form we may find the plural: *lease of ****promises***. We detected a case of a possible contamination: ****îngenunchia* instead of *îngenunchea*[3] (improper suffixation); improper suffixation was detected in an English term as well: *sovereignity* (Lozinschi, 2008).

Prepositions as headwords are questionable in a specialized dictionary, especially when multiword expressions begin with a preposition: ***în** baza ordinului* ('under the order'), as readers will search for the expression under *bază* or *ordin*.

[3] There is a tendency of phonetic opening in Romanian.

Combined words may cause misunderstanding as well, since instead of *footwear* we found **footware* (cf. *hardware, software*), which may not be a typographical error as it appears twice (Lozinschi, 2008). However, this type of error already leads us to another category.

3.1.4. Linguistic errors

We had a prediction at the beginning of the research that the more languages are involved, the more mistakes are to be expected and our anticipation seems to be right in the following cases:

- Romanian influence upon English (Lozinschi, 2008): *draw up a* ***raport** instead of ***report***, as the Romanian *raport* means *report*. Similarly, we have **set up* ***o*** *company* instead of *set up* ***a*** *company* (Voroniuc, 1999), where the Romanian *o* is the equivalent of the English *a* (indefinite article). Another example is **negociations* (< Ro. *negociere*).
- English influence upon Romanian (Botezat, 2011): *process* (English) instead of *proces* (Romanian) under the headword *încetare*.
- French influence upon English (Lozinschi, 2008): *assu**r**age* (typographical error, extra letter) in French is a term used in alpinism, and in our case it is used instead of *assuage* (or the document writer was set to French); similarly, *revigoration* (French) is used instead of *reinvigoration*. We may even call these errors false friends, but they still remain errors.

3.1.5. Formatting mistakes

Under formatting mistakes we refer to any deviation from standard reference for the entries, namely symbols, abbreviations, punctuation and layout; deviations from these standards may be either visually bothering or completely wrong. For instance, the tilde sign (~) replacing the headword is a generally accepted norm, which is overlooked in Hanga and Calciu's dictionary. In the same dictionary more than one type of abbreviation is used for the same thing, for instance the difference between American English and British English is marked with *US, SUA* (the Romanian equivalent for USA), *U.S.A.* and *amer.* (American), and *UK, în Anglia* ('in England'), respectively.

The worst type of mistake within the category definitely goes to the alphabetical order. Talking about dictionaries, it is a must to have a 100% correct alphabetical order, although this basic rule was not taken into consideration in the case of at least two dictionaries. Hanga and Calciu's dictionary contains *cadastru* before *cabotaj*, whereas Dumitrescu's dictionary has a problem with the alphabetical order when the Romanian diacriticals *ș* and *ț* are involved. Thus page 122 (Dumitrescu, 2009) contains *înș**t**iințare–înț**e**les* terms (21 in number) first, followed by *în**a**inta–în**s**cris* terms (5 pages), then *înș**e**lăciune* ('racketeering') appears with 3 other terms, to be followed by *în**s**trăin**ă**ri* and *înstrăin**a*** ... (*ă* before *a*). Thus it is not only *ș* and *ț* is problematic, but the other three diacritically marked letter as well (*ă, î, â*). As a result, we have serious doubts that the reader will find the term, even if it included in the dictionary.

3.2. The importance of term base management

The proper knowledge of the term base manager presupposes the knowledge of the previously presented error types in order to avoid them when creating a term base. Term bases (TB) are electronic databases and are typically accessed via computers (online or with the help

of specialized software). In our case, creating a term base for legal terms in Romanian and English means at least three things:

- enough input from various sources (more Romanian–English law dictionaries);
- selection of entries belonging to law (even if the separation of legal and economic terms in Romanian–English context is not an aim) and not general terms (body parts, numbers, countries, regionalisms, etc.);
- a proper care for the words containing Romanian diacritical marks (letters *ă, â, î, ș, ț*), as they may not appear in the database adequately or in the wrong alphabetical order.

After the term base is created, our aim is to be used when translating legal documents from/into Romanian and English.

4. Creating a legal database

After having purchased these dictionaries, the major aim was to cross-examine them in an effective way and trying to detect their flaws and correct them. Thanks to a POSDRU project at Petru Maior University in Tg.-Mureș we have embarked upon creating a common database from all these sources. In case we start with the most comprehensive dictionary, the others are much easier to include in the database. So we considered that we should start with Lozinschi's dictionary, which proved to contain nearly 100,000 entries, although many of them do not belong to the legal terminology. The other six dictionaries were selected due to other factors (notoriety, foreign authors, etc.). In our view, present-day translators should be able to create databases compatible with computer-assisted translation software. The most widespread CAT-tools are compatible with each other – at least, to a certain extent[4]. Thus creating a term base in *csv* format should be compatible with various, even cross-platform CAT tools. A *csv* format may be easily obtained after having created a *Microsoft Excel* file (*xls* or *xlsx* format) with two columns (in our case Romanian and English), then converted into *csv* format. This type of format can be easily used as an external term base for translation environments. After having collected data from the seven dictionaries, we obtained the following number of entries for letter *Î*:

Dictionary	**Nr. of entries**
Lozinschi, 2008, pp. 263–304	7,410
Hanga & Calciu, 2009, pp. 90–93	199
Lister & Veth, 2010, pp. 442–451	1,155
Dumitrescu, 2009, pp. 120–128	788
Botezat, 2011, pp. 88–92	314
Voroniuc, 1999, pp. 60–63	385
Voroniuc, 2011, pp. 284–288	445
TOTAL	10,696

Table 3. Number of entries form seven dictionaries letter Î

[4] A list of notable CAT tools may be checked here: http://en.wikipedia.org/wiki/Computer-assisted_translation, 04. 10. 2014.

One should note that when a TB is created, one entry means that one word/expression in the source language "equals" one word/expression in the target language, so if we have three translations for *încărcătură*, we will have three entries: *încărcătură–freight, încărcătură–cargo* and *încărcătură–load*. We can observe that there is an enormous difference in number regarding these dictionaries; basically, Lozinschi's dictionary contains more than 37 times the number of entries to be found in Hanga and Calciu's dictionary, or it contains more than twice the number of all entries compared to the other six dictionaries. Partially, it is more detailed than the others, but it also contains many entries not belonging to the legal terminology.

5. Conclusions

Professional translators already signaled on www.proz.com that the quality of a Romanian–English law dictionary is below expectations[5]. However, the others contain extremely many errors in many respects detailed in this article. What is worse is that we cannot say that the quality is improving with the advances in technology and the passing of time: Dumitrescu's dictionary cannot follow the very basic alphabetical order of entries, whereas Voroniuc's 2011 dictionary republishes all the previous errors from 1999 and even manages to add further ones: *întrerupere* as headword is added, containing **negociations* (Romanian influence, < Ro. *negociere*). Similarly, a new entry is *înțelegere*, containing **convenant* instead of *covenant* (probably Romanian influence, e.g. *convenabil*). The only improvement is the replacement of *î* into *â* in the middle of the words due to the changed Romanian spelling rules (e.g. *dobândă* replacing *dobîndă*).

Having seen these errors it is no wonder that there are experts who say that at present we can witness a certain vulgarization of science/language, although many of them might be easily solved, for instance setting the correct typing language in an office document (*Microsoft Office, Libre Office*) and the spell-checker is adjusted to the desired language. Thus we can obtain rather error-free results (Imre, 2013). This is why we consider it disturbing that Hanga & Calciu's dictionary is at its fifth re-checked and completed edition[6]. A well-founded question is, what was the first edition like?

And finally, the choice of UK or US English may be solved very easily: the preface of the dictionary should specify whether the author wishes to follow one or the other terminology/spelling and only the entries in the other language should be marked. This is not systematically followed in any of dictionaries involved so far.

Acknowledgements : The author wishes to thank the invaluable support regarding the Romanian data and the term base to Doina Butiurcă, Beatrice Standavid and Blanka Barabás.
The research presented in this paper was supported by the European Social Fund under the responsibility of the Managing Authority for the Sectoral Operational Programme for Human Resources Development (*Sistem integrat de îmbunătăţire a calităţii cercetării doctorale şi postdoctorale din România şi de promovare a rolului ştiinţei în societate*), as part of the grant POSDRU/159/1.5/S/133652.

[5] Here are only two links to prove it: http://www.proz.com/kudoz/romanian_to_english/law_contracts/3490819-prepusi.html, http://www.proz.com/kudoz/english_to_romanian/law_patents/479796-invalidity.html. 04. 10. 2014.
[6] In Romanian: "Ediția a V-a revăzută și adăugită", interior cover/first page.

References and dictionaries

Bantaș, A., & Năstăsescu, V. (2000). *Dicționar economic român-englez*. București: Editura Niculescu SRL.
Barber, B. R. (1992, March). Jihad vs. McWorld. *The Atlantic Monthly*, (3), 53–63.
Bloom, D. (2012, September 30). Spell checkers developing "atomic typo" capabilities. Retrieved October 5, 2014, from http://www.chinapost.com.tw/commentary/the-china-post/special-to-the-china-post/2012/09/30/356026/Spell-checkers.htm
Botezat, O. (2011). *Dicționar juridic român-englez / englez-român* (2nd ed.). București: C.H. Beck.
Dumitrescu, D. (2009). *Dicționar juridic român-englez*. București: Akademos Art.
Gouadec, D. (2007). *Translation as a Profession*. John Benjamins Publishing.
Hanga, V., & Calciu, R. (2009). *Dicționar juridic englez-român și român-englez*. București: Lumina Lex.
Imre, A. (2013). *Traps of Translation*. Brașov: Editura Universității "Transilvania."
Ionescu-Cruțan, N. (2006). *Dicționar economic englez-român, român-englez*. București: Teora.
Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press.
Lister, R., & Veth, K. (2010). *Dicționar juridic englez-român / român-englez*. (R. Dinulescu, Trans.). București: Niculescu.
Lozinschi, S. (2008). *Dicționar juridic Român - Englez*. București: Editura Smaranda.
Năstăsescu, V. (2006). *Dicționar economic englez-român / român-englez*. București: Niculescu.
Snell-Hornby, M. (2006). *The Turns of Translation Studies: New Paradigms Or Shifting Viewpoints?*. John Benjamins Publishing.
Trumble, W. R., & Stevenson, A. (Eds.). (2002). *Shorter Oxford English Dictionary* (Fifth Edition.). OUP Oxford.
Typographical error. (2014, October 3). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Typographical_error&oldid=628064781
Voroniuc, A. (1999). *Dicționar de termeni economici și juridici (român-englez)*. Iași: Institutul European.
Voroniuc, A. (2011). *Dicționar englez-român / român-englez de termeni economici și juridici*. Iași: Polirom.