A. Gelbukh, I. Bolshakov. *Internet, a* true *friend of the translator*. International Journal of Translation, International Journal of Translation, ISSN 0970-9819, 2003, to appear.

Internet, a *True* Friend of the Translator*

Alexander Gelbukh and Igor A. Bolshakov

Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Mexico City, Mexico
{gelbukh,igor}@cic.ipn.mx; www.Gelbukh.com

Abstract. In practice of manual translation, most problems can be reduced to the word choice problem and to the problem of understanding of an unknown foreign word. We show how the lexical richness of Internet can be used to semi-automatically solve these problems. Namely, Internet search engines provide useful statistics on word usage frequencies, word combinability, and word usage context. However, the use of this statistics is not straightforward and requires some precautions, which we discuss in the paper. For the same reason, purely automatic application of these techniques seems still impractical, though we show that useful tools for their semi-automatic application can be developed.

1 Introduction

It is well known that to translate is to understand the text in one language and to tell the obtained idea in the other language. This is certainly true for human translation and, somewhat arguably, for automatic translation.

Humans and computers perform differently these two tasks. For a good human translator, understanding is usually no problem, while formulating the other's ideas in a fluid and idiomatic manner is difficult even in one's mother tongue. Of course, translation into one's non-native language presents much more problems while composing the target text. Though for a computer the understanding is currently still a

* Work done under partial support of Mexican Government (CONACyT, SNI), CGEPI-IPN (Mexico), and RITOS-2.

bigger problem, it also faces all those difficulties that a human writer faces in composing text in his or her non-native language (indeed, computer has no native language).

Accordingly, in this paper we will mostly concentrate on the task of composition (or, as it is called in computer-related context, *generation*) of text in one's non-native language.[1] What is more, we will mostly concentrate on human translation, though the techniques we will discuss can be applied to improve machine translation programs too.

Text generation involves several difficulties, such as discourse planning, rhetorical structure planning and realization, choosing correct syntactic constructions, etc. For an experienced human translator these tasks are usually not very difficult (for a computer they are). What is the real pain in the neck in translation process is word choice problem: Which word to use to translate a given word or an idea in a given context?

Word choice problem appears because of several reasons, for example:

- Homonymy and polysemy. The source word can have different meanings, which are usually translated into different words in the target language. How would you translate, say, *bill* into Spanish? The possible translations are *cuenta* 'check', *factura* 'invoice', *billete* 'banknote', *propuesta* 'law', *pico* 'beak', *hacha* 'hoe', etc. Thus, knowing from the high school that "*bill* is translated as *cuenta*" is in fact no help (if not harm) in translation.
- Synonymy. The same meaning can be expressed with different words in the target language. Say, *high* and *tall* both express the same meaning 'having big altitude' and thus are both pretty good translations for Spanish *alto* (in this meaning). However, it is well known that synonyms are not mutually interchangeable in the text. For example, one has to say *tall man* and *high degree*, though one can say both *high hat* and *tall hat* (while in Spanish the same word *alto* is used in all three cases). Again, the school rule "*alto* is translated as *high* or *tall*" does not work.

[1] In the examples, we will usually assume translating from one's native language into English and illustrate the translation difficulties on English constructions, to make them understandable to the reader. As the other language we will mostly use Spanish, though no knowledge of Spanish is expected from the reader.

- Imperfections in dictionaries or wrong guesses of the translator; false cognates (called also false friends—from the French *faux amis*—or *false friends of the translator*). Some words are not translated into another language as their outer shape might suggest. For example, Spanish *asistir* 'to attend' (*asistir congreso* 'to attend conference') does not mean *assist* as one might expect, cf. *admitir* 'to admit', *permitir* 'to permit', *omitir* 'to omit', etc.; Spanish *idioma* 'language' (*idioma inglés* 'English language') does not mean *idiom*. On the one hand, such incorrect translations can sometimes be given in (bad) dictionaries due to negligence of their authors. On the other hand, due to language changes some false friends can with time become loan words (or meanings) in the language in question, while for a long time they will still be absent in the classical dictionaries. Say, in Mexican Spanish the words *checkar* 'to check', *accesar* 'to access' are commonly understandable and used, though they are still absent even from Microsoft Word spell-checker dictionary since traditional dictionaries consider them false friends, having the traditional expressions *verificar* 'to check' and *tener acceso* 'to access'.
- Lexical functions and terminological usage. In many cases a word is used to express an idea having nothing to do with what dictionaries state as its meaning. Indeed, in the expressions such as *pay attention*, *give birth*, *high wind*, *strong drink* one of the words is not used in its "normal" meaning and, no surprise, is not translated by the "normal" counterparts that bilingual dictionaries list. For example, *pay* (*attention*) is not translated into Spanish as **pagar* (*atención*) 'pay' but as *prestar* (*atención*) 'loan' (note that the literal translation of this absolutely correct Spanish expression would be **loan attention*, which is not understandable in English).
- Idiomatic usage. A whole expression can have a meaning not composed of the words it consists of. For example, *hot dog* (fast food) is not a dog that is hot. Such expressions can sometimes be translated literally into some languages, but usually they are not. For example, though in modern Mexican Spanish the expression *perr(it)o caliente* '(little) dog that is hot' is already understandable colloquially as *hot dog*, it cannot still be used formally in this meaning.

In the first two cases (homonymy and synonymy), the possible translation variants are given in the dictionaries. However, most of the dictionaries are oriented to reading foreign language texts rather than writ-

ing them. Thus, they usually do not give the information sufficient to choose one of translation variants in text composition.

In the third case (false friends) the correct translation is also present in the dictionary, though the translator does not check it. An automatic or semi-automatic procedure to detect such errors is especially desirable.

As to the last two cases (lexical functions and idioms), in most cases dictionaries do not help; the translations are given in the dictionaries only for the most common cases.

Thus, in the first three cases the word choice problem in composing text in the foreign language is reduced to forming several possible hypotheses (with dictionary's help) and choosing between them. In this paper we will mostly deal with such cases, not with lexical functions or idioms. Our discussion is also applicable to other cases when a choice out of a small number of possible hypotheses is to be done, such as in preposition choice: for example, is Spanish *en el congreso* translated as **in the conference*, **on the conference*, or *at the conference*?

The point of our paper is that most of the recommendations given below can be applied manually using a standard web search engine such as Google. However, we are interested in the possibility of their automatic or semi-automatic application either in a tool helping human translator or in an automatic translation program.

In Section 2 we present the idea of using Internet to verify the translation hypotheses. In Sections 3, 4, 5, and 6 we discuss how the style of the document, the word meaning, the context of the word, and morphological inflection can be taken into account and what precautions are to be taken in obtaining the corresponding statistics from an Internet search engine. In Section 7 we consider how Internet helps in understanding (rather than composing) a text in the foreign language. In Section 8 we give a more precise formula for calculation the statistics of word combinability. In Section 9 we discuss semi-automatic tools that simplify the use of the presented techniques in the translator's practice. In Section 10 we explain why purely automatic application of our techniques in machine translation is difficult (though not impossible), and give an idea of a semi-automatic tool for detection and correction of the false friends.

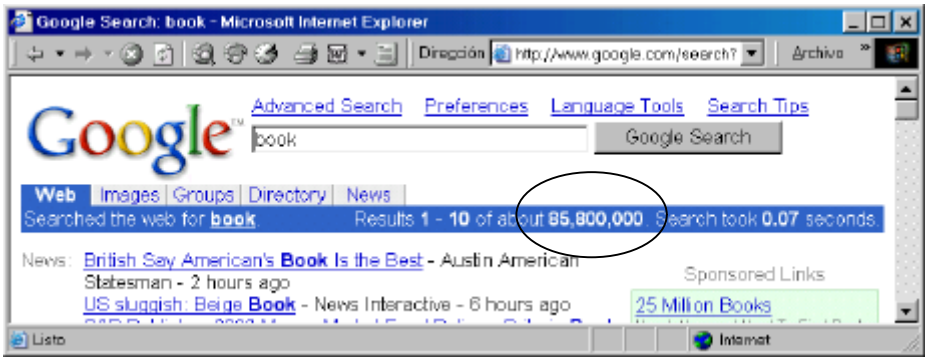

Fig. 1. Google indicates the number of pages found at the top of the result page.

2 Internet and Word Choice Problem

The idea of using Internet to verify translation hypotheses is very simple: what is more common is more correct, and Internet helps us to count usage frequencies (most Internet search engines indicate the number of documents found for the query, see Fig. 1, Fig. 2).

For example, for the meaning expressed with the Spanish word *hipopótamo* the dictionary we used [20] gives as possible translations *river*(-)*horse* and *hippopotamus*. Which word would you use in your translation (remember that we assume that English is not your native language)? Google search gives:[2]

Query	Number of documents
"*river horse*"	7060
riverhorse	1550
hippopotamus	87000

This suggests that people are more familiar with the word *hippopotamus*, which is perhaps the best candidate for translation.

[2] Google and AltaVista seem not to have any means of specifying in the query the difference between "*river-horse*" and "*river horse.*" In fact, both queries match any document where the word *river* precedes the word *horse*, e.g.: "*the trail that leads to the (Lewis River) horse camp*" (Google).

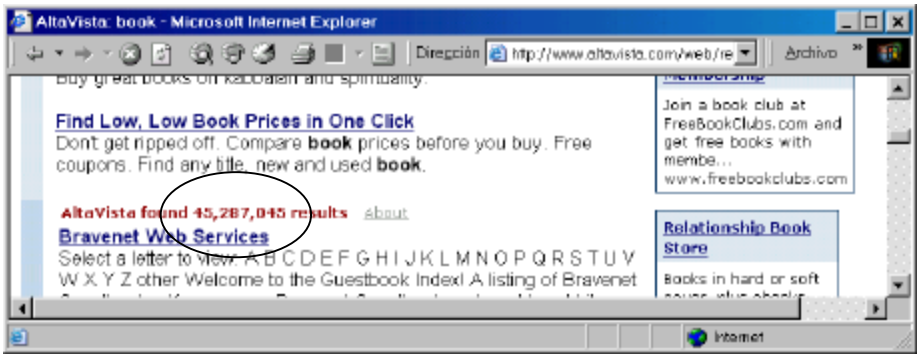

Fig. 2. AltaVista indicates the number of pages found after the advertisement links in the list.

An important precaution is to make sure your statistics is valid for the given language. For example, while the expression *river horse* is most likely to be found in English texts, the word *hippopotamus* can also be found in the texts in Latin, Indonesian, Slovak, and Tagalog [15], which might lead to unfair statistics. Of course, these languages perhaps do not present a great the problem, but here is a better example: for Spanish *valor* the dictionary gives *valor* and *bravery* among other variants; which one is more common? In the following table we show the results obtained when Google was configured to search in English language pages only (see below):

Query	Any language	English only
valor	1440000	224000
bravery	374000	202000

This shows that *valor* has nearly the same frequency in English as *bravery*, though in Internet in general it is three times more frequent due to coincidence with the Spanish word *valor* 'value'. Thus, you should configure Google to search only English (in our case) pages. This can be done manually in the Language Tools page. Alternatively (and also in automatic applications), you can change the language settings in the query URL in the browser's address bar. Say, if you are looking at Google's results for *valor*, you would see in the address bar something like this:

```
http://www.google.com/search?lr=&cr=&q=value&hl=en&ie=ISO
-8859-1&safe=off
```

and to instruct Google to search only in English pages, you should change the URL as follows (we found this by manually changing the settings in the Language Tools page and watching the changes in the URL):

```
http://www.google.com/search?lr=lang_en&cr=&q=value&hl=en
&ie=ISO-8859-1&safe=off
```

What is more, make sure you search in the pages with good language use — basically, in the pages written by native speakers. This is especially important for the languages used for international communication, such as English. To assure this, you can specify a country in your query: say, to search only in British or United States sites. You can specify this in the query or in the options for your search engine. For example, for Google this can be specified in the Language Tools page or in the query URL as follows (*gasoline*, English, Great Britain):

```
http://www.google.com/search?lr=lang_en&cr=countryUK%7Cco
untryGB&q=gasoline&hl=en&ie=ISO-8859-1&safe=off
```

The following table shows that the expression "*talk on the conference*" is uncommon but still used in the world (3% of cases). However, it is nearly impossible in Great Britain (1% of cases):

Query	Any country	Great Britain
"*talk on the conference*"	64	2
"*talk at the conference*"	1990	134

This is mostly due to bad language use by non-native speakers, e.g.: "*This paper contains the subject of my talk on the Conference in Twente in December 2000*" (Russian site), while the only two British texts retrieved actually contained the query string by accident: (1) "*An attendee can now talk on the conference call*"; (2) "*Barry Took will liven delegates up with his amusing talk on the conference theme.*"

Of course, with this technique you risk to obtain the results for a specific language variant (British English in our case). On the other hand, this can be exactly what you need. E.g., is *petrol* as common as *gasoline* or *petroleum* in the United States? And in Great Britain?

Query	United States	Great Britain
petroleum	1240000	128000
petrol	67000	82000
gasoline	925000	23100

3 Document Style

If the problem were as simple as is described in the previous section, the dictionaries would always give only one translation variant for each word. In reality, several factors affect the choice, first of all, style, word meaning, and context.

First, let us consider the style. Is it true that in scientific style, *hippopotamus* is more appropriate, while in the belles-lettres style *river-horse* is preferable? There are two ways to evaluate this:

- Refine your search. Say, you are interested in scientific style. Try adding to the query those words that would tend to give scientific (or belles-lettres) style documents, for example:[3]

	physiology	*scared*
"*river horse*"	28	234
hippopotamus	840	2360

  Note that even though the word *hippopotamus* is more common in both styles, *river horse* is still more acceptable in belles-lettres style (10% of usage cases) than in scientific style (3%).
- Count the proportion of the desired documents, i.e. take into account scaling factors. In many cases it is difficult or unreliable to invent a query that would guarantee a certain style and would not lead to skewed statistics. You can workaround this problem by evaluating the share of the desired documents in the search results. For each query, take a random sample of the returned documents and classify them by styles. Say, you found that only each 30th document in the results for "*river-horse*" was of scientific style. Then you should replace the number 7060 in the table at the beginning of Section 2 for 7060 / 30 = 235. Similarly, if each 10th document found for *hippopotamus* was of scientific style, for the corresponding row of the table from Section 2 we have 87000 / 10 = 8700.

[3] The figures stand for the queries *physiology AND "river horse"*, etc. For simplicity, we omit the parts of the query specifying the language and country; in real application they should be set as explained above. Make sure to use correct syntax to express AND in the search engine you use. For example, for some search engines the syntax would be "+*physiology* +*hippopotamus*".

We recommend applying both techniques: do refine your search if you can, and then analyze a sample of the results to verify that your query retrieves only the documents of the desired style.

When counting, be sure to take a representative sample of the documents. The search engines have their own policy of ordering documents, which is usually not appropriate for statistical research [16]. We can recommend considering, say, each 10^{th} or each 100^{th} document from the search results, rather than the first several documents. In Google, this can be done by looking at the first document in each page of the search results, using "Next" link to go to the next page or by directly specifying the desired number in the query URL:

```
http://www.google.com/search?q=%22river-
horse%22&hl=en&lr=&cr=&ie=UTF-8&safe=off&start=600&sa=N
```

to go to the page containing the document number 600. Note that different search engines have different limitations on the maximum number of documents they can show to the user: Google currently show up to 1000 documents, AltaVista up to 200.

4 Word Senses

In addition to Spanish *hipopótamo*, we consulted other dictionaries [18, 1] for the Russian word *begemot* having exactly the same meaning.[4] As possible translations we got *river*(-)*horse*, *behemot*(*h*), *sea*(-)*cow*, and *hippopotamus*, with the following frequencies:

Query	Number of documents
"*river horse*"	7060
riverhorse	1550
"*sea cow*"	11200
seacow	2230
behemot	6530
behemoth	180000
hippopotamus	87000

This suggests that the most common translation is *behemoth*. However, one of the dictionaries [1] marks this word as related to Bible, which suggests that the word might have several meanings. Also, the

[4] We can assure this since we are Russian native speakers.

inverse translation of the word *sea-cow* into Spanish gives several other words, e.g., *morsa* ‘walrus’, which at least suggests that this word also is homonymous. In fact, it is much more common that a word is homonymous than not, e.g., *bill* (see Section 1); *well* ‘deep hole’ vs. ‘in a good manner’, etc.

Therefore, you should always check if the word in question is used in the retrieved documents in the meaning you need. This can be done with the same two techniques:

- Refine your search to likely retrieve the documents containing the word in the desired sense. For example, the animal referred to by the Spanish *hipopótamo* and Russian *begemot* lives in Nile, while walrus does not. We get:

Query	Number of documents
Nile AND “*river horse*”	545
Nile AND “*sea cow*”	295
Nile AND *behemoth*	3620
Nile AND *hippopotamus*	5450

- Analyze a random sample of the documents, estimating the share of the documents in which the word is used in the desired meaning.

Again, both techniques should be applied. In the previous table, the second best choice is still *behemoth*. However, when we analyzed a sample of the documents retrieved with the query *Nile* AND *behemoth*, we suspected that this word does not refer to an animal, at least not to a real animal. So we tried another search:

Query	Number of documents
zoo AND “*river horse*”	585
zoo AND “*sea cow*”	2140
zoo AND *behemoth*	2280
zoo AND *hippopotamus*	9810

In this experiment, *sea cow* looks like a possible candidate. To resolve the doubts, we tried another search:

Query	Number of documents
“*river horse attacked*”	1
“*sea cow attacked*”	0
“hippopotamus attacked”	22

These experiments show that a dangerous animal living in Nile is most likely to be referred to as *hippopotamus* and not as *behemoth* or *sea cow*, even if the dictionaries give these translation variants (in fact, *behemoth* proved to be a clear example of a false friend for Russian *begemot*, and the variant *sea cow* is more likely an error in the dictionary).

5 Collocations

In a sentence a word modifies, and/or is modified by, other words, forming word combinations (called also collocations). A collocation is defined as syntactically related (in the sense of dependency grammar [19]) pair of content words connected through a possible chain of functional words, e.g., *dig* (*a*) *well*, *fetch* (*water*) *from* (*a*) *well*, *water in* (a) *well*, etc. [3, 4]. Fortunately, specifying in the query the context words with which the word in question forms a collocation frequently resolves other problems such as stylistic nuances and word sense ambiguity. For example, in the expression *fetch water from a well* the word *well* is most probably a noun and not an adverb (but: *fetch water from a well positioned can*).

However, most important property of collocations is that words with the same basic meaning are frequently selective as to with which other words they can form collocations. For example, though *river horse* can refer to the animal, it is not used to refer to its skin:

Query	Number of documents
"*river horse skin*"	0
"*sea cow skin*"	8
"*behemoth skin*"	24
"hippopotamus skin"	240

Another example: Spanish *alto* is translated as *tall* or *high*, but these words have different combinability with other words:

	school	*mountain*	*man*	*tree*
tall	274	6170	116000	33000
high	5120000	173000	5770	10800

Similarly selective are prepositions. Spanish preposition *en* can be translated as *on*, *in*, *at*; which is the correct translation in the phrase

encontrarse en el congreso ‘meet at the conference’? The first attempt gives:

Query	Number of documents
“*meet on the conference*”	19
“*meet in the conference*”	1400
“*meet at the conference*”	651

which apparently suggests that the correct variant is *meet in the conference*. However, analyzing the documents retrieved for this query, we found that a frequent pattern in them is “*meet in the conference room*”. So we tried another Google search: *"meet in the conference" -"meet in the conference room"* (which stands for the documents that contain *meet in the conference* but do not contain *meet in the conference room*), which gave only 325 documents (of them, many were *meet in the conference hall*, etc). With this, the best choice resulted to be *meet at the conference*.

6 Morphology: Which Search Engine is Better?

A complication similar to word sense noise is caused by morphological inflection. If you count the documents where the given word appears in any morphological form (*to leave*, *leaves*, *left*, *leaving*), you risk to count in another word that by accident coincides with the given one in one of its inflectional forms, cf. *to the left from the window*, *the leaves of the tree*. Note that this problem is much more frequent in languages with richer morphology (in Spanish: *como* ‘I eat’ or ‘as’, *nada* ‘swims’ or ‘nothing’, *haz* ‘bundle’ or ‘do’, etc.).

The problem is that if you only count the main form of the word (*leave*, but not *left* nor *leaves*), you risk missing many relevant occurrences. To make a decision, you should search for all morphological forms (if your search engine supports this, such as Russian search engine Yandex [22]) and analyze the effect on a random sample. Basing on this analysis, you can correct the resulting figure. Alternatively, if you found that only a rarely used form of the word causes the problem, you can use morphological inflection but explicitly exclude the problematic form from the search.

The major search engines, however, do not support morphological inflection. Then, to search for all forms of a word, you need to explicitly list them in your query, e.g., *leave OR leaving OR leaves OR left*.

However, we found that search engines do not give correct figures for the number of documents found when OR operator is used. Especially absurd results are given by Google, as can be seen from the following table, where Google reports for *leaves OR left* 10 times less documents than for *left*, which is logically impossible:

Query	AltaVista	Google
leave	16278127	26200000
leaves	5755935	9190000
left	28336174	51900000
leaves OR left	32159277	9080000
leave OR leaving	16364865	7690000
leaving OR leaves OR left	30534639	9930000
leave OR leaving OR leaves OR left	36960363	8800000
leave AND NOT left	17878438	6170000
leave AND NOT left AND NOT leaves	12446331	6650000

(Here we give AltaVista syntax.) Thus, our recommendation is not to use Google (or use it with great precaution) for complex logical queries, at least those containing OR.

7 The Inverse Problem: Guessing Word Meaning

In the previous sections we were mostly concerned with the word choice problem: which word is to be chosen in composition in a foreign language? Now consider the inverse problem: given a word in the foreign text, what is its meaning and usage? In fact, word choice problem involves the same question: what are the meaning and usage nuances of each translation variant?

Internet helps in resolving this problem by providing a huge amount of usage examples. The general procedure is the same as above: type a word in a search engine's interface, analyze a random selection of the retrieved documents (do not look at the very first page of the results but instead go to, say, 10th page, see Section 3), guess the meaning, refine your search query to verify your hypothesis, and repeat these steps until you have a clear idea of the meaning.

Let us return to the first table from Section 4, which suggests that *behemoth* is a very common word and thus must be a good translation for Russian *begemot* 'hippopotamus.' A quick analysis of the returned

pages gave a lot of garbage such as company, product, or musical band names, song titles, game characters, and some unclear contexts. A query "*behemoth -title -disc -metal -team -song -game -little*" (*behemoth AND NOT title AND NOT disc* etc.) over English pages reduced the results to 55000 pages. Here are excerpts of the found pages:

...Return of the Soviet ***behemoth****. The world's largest aircraft, the Antonov-225, returns to the skies at the Paris Air Show...*
...What is it that makes them behave as if monopolistic bullying is an inherent right for the biggest, baddest damned ***behemoth*** *ever to roam the face of the information economy?...*
...A Bacteria ***behemoth****. The discovery and characterization of* Epulopiscium fisheloni*, the world's largest bacteria, has created a host of problems for microbiologists. First off, the* ***behemoth*** *size of the organism questions the surface-to-volume ratio necessary for the survival of cells...*

No surprise, these examples (and other examples we found) agree with the definition from [**¡Error! No se encuentra el origen de la referencia.**]: *Noun.* 1. *Something enormous in size or power.* 2. *often Behemoth: A huge animal, possibly the hippopotamus, described in the Bible.*

However, this technique can be more useful for some other words that do not appear in dictionaries—such as *Google*—or for local words or rare or slang words in languages for which no good dictionaries are available, e.g., Mexican Spanish *chilango* 'one from Mexico City' (what is more, such words can be often found in Internet dictionaries like [9]).

8 More Precise Collocation Statistics

In Section 5 we discussed the numbers of occurrences of word pairs, called by different authors collocations or bigrams [**¡Error! No se encuentra el origen de la referencia.**, **¡Error! No se encuentra el origen de la referencia.**]. However, such expressions can be formed by accident because of coincidence of the words unrelated in the linguistic structure of the text, as we have seen in Section 2 ("*the (Lewis* ***River) horse*** *camp*") or Section 5 ("***in*** *the* ***conference*** *room*"). For a better estimate of the number of real (linguistic) collocations, let us roughly estimate the number of such accidental (false) collocations in a large corpus (such as Internet).

To simplify calculations, we will (naï vely) suppose that the words appear in the text statistically independently. Let the given search engine has indexed N pages, of W words each on average; then there are roughly NW pairs of words (one immediately after another) in all the indexed texts. Consider two words, w_1 and w_2, with the corresponding number of occurrences n_1 and n_2 in the indexed texts. Then the expected number of the pairs consisting exactly of the words w_1 and then w_2 is $(n_1n_2)/(NW)^2$, while the expected number of pages where they would occur is $(n_1n_2)/(NW)$. This number is to be subtracted from the figure given by the search engine to approximately find the number of non-accidental (real) collocations.[5]

For application of this formula, the values of N and W are to be found. We do not have any reliable figure to W, so we will just consider it to be 1000. As to N, this proves to be rather difficult [6]. One approximation is what the search engine itself reports. For example, Google's main page states that Google have indexed (at the moment of writing this paper) $N = 3083324652$ pages. However, this number includes the pages in all languages, so that for two English words, the independence hypothesis does not hold. Perhaps a better way to determine the number of indexed pages for a given language is to use a query with a very common word, such as *the*, or a disjunction of several very common words (which, as we have mentioned in Section 6, does not work with Google). For a strange reason, Google shows nearly the same figure for the query *the*: 3,220,000,000. However, if we restrict the search by English pages only, for *the* we get $N = 13{,}800{,}000$.

Let us apply this idea for the pairs "*eat book*" and "*obsolete book*". Google search in English pages only gives:

	obsolete	*eat*	Word alone
book	213	329	8060000
Word alone	222000	3070000	

which suggests that *eat book* is a better collocation than *obsolete book*. However, subtracting the result given by the above formula, we get:

	obsolete	*eat*
book	213−129=+84	329−1793=−1464

[5] A related statistical concept is so-called mutual information, however, we do not discuss it here since the results obtained with this measure are more difficult to interpret by a non-specialist.

This indicates that *obsolete book* is a much better collocation than *eat book*. Indeed, typical contexts for "*eat book*" were "*What Would Jesus Eat? Book review*"; "*Those prehistoric looking critters eat book glue and spread like wildfire*"; "*This was a grrrr-eat book teaching me all about bats*", while for "*obsolete book*" a typical context is "*the first edition is an obsolete book...*"

The negative result in the table can indicate either that the linguistic properties of these two words actually prevent them from being located next to each other even by chance, or that our choice of the parameters *N* and *W* is not quite correct.

9 Semi-Automatic Tools

The discussion in the previous section suggests that it is a good idea to develop a semi-automatic tool that would apply automatically all necessary calculations, such as those described in Section 8. Though we are not aware of such a tool, we believe that it is quite easy to develop some helpful routines of this kind in any programming language supporting automatic web page download, such as Perl. The search engines can be consulted through URLs similar to those discussed in Section 2.

More traditional word usage statistics tools use text corpora instead of the full Internet contents. However, currently corpus statistics tools seem to provide the results in the form better suited for lexicographic research. An example is concordance tools [11], some of which are available through Internet interface (though they use traditional corpora as database). Here is an example output of the concordance tool [10] for the queries *behemoth* and *hippopotamus*, which allows to guess the meaning and combinatorial properties of these two words:

will only slow down this industrial **behemoth**. It won't stop it. In fact, if
might not take kindly to a C&W-BT **behemoth**. Thanks to the strong stance taken
UK visit. A shifting, throbbing **behemoth**, there are also two new mixes from
a three-thousand-member political **behemoth**, whose hulking presence in
in less than two decades into a **behemoth** with $ 22 billion in
in the 1970s, was a sixty-seven-ton **behemoth** with thermal sites that permitted

An incident occurred in which a **hippopotamus** attacked a kayak from
Nile, all they come up with is **hippopotamus** crap? [p] NICHOLAS LEZARD [p]
you might write horse and **hippopotamus**; dog and donkey; kid and
a situation involving a **hippopotamus** in Zimbabwe. [p] A small group

drowning, has saved the hippopotamus in the Kruger Park and other
a narrow escape when coming on a hippopotamus with her calf in mid-river. The

One can see that this form of output is more convenient than analyzing the web pages one by one. However, the traditional corpora used in the tools like [10] are much more limited in size: the largest corpora are hundred thousand times smaller than Internet, so that they just do not provide enough statistics for the vast majority of the words in the dictionary of a given language, not speaking of word combinations.

There are attempts to overcome this problem combining the advantages of the huge corpus provided by Internet with the convenience of the traditional locally stored corpora. For example, representative corpora [**¡Error! No se encuentra el origen de la referencia.**] contain much of the same statistical richness that Internet provides, and still are stored locally and can be used by traditional concordance tools.

Finally, yet another important tool provided by Internet—though having no relation with its textual contents—is web-based interfaces to plenty of dictionaries and wordlists for nearly all languages, a treasure no library in the world has [**¡Error! No se encuentra el origen de la referencia.**, 9].

10 Challenges and Perspectives of Automatic Application

Can we, however, go beyond semi-automatic tools that merely speed up routine manual operations? Can Internet statistics be applied in a fully automatic manner to machine translation?

As we have seen, the direct application of the techniques discussed in this paper requires certain ingenuity, as well as linguistic operations—such as word sense disambiguation or style recognition—that cannot still be reliably performed in a fully automatic way. Though the enormous potential of Internet lexical richness will eventually give rise to a new generation of purely automatic linguistic applications [8], we believe that development of semi-automatic tools that leave the difficult decisions on the human expert is, for the moment, a promising direction in supporting the work of practical translators.

In some cases the program still can purely automatically find the necessary word and suggest it to the human translator. An example of such an application is detection and correction of the false friends. Indeed, a false friend is a word for which the translator thinks that he or

she knows the correct translation, while in fact they do not. The problem here is that the translator would not consult any dictionary to translate such a word—just because his or her confidence. To detect the error, one needs to verify nearly each word in the text the translator composes. Obviously, it is desirable to implement such verification automatically.

Our idea of such a verification tool (whose implementation will be a topic of our future works) is based on the algorithm developed by Hirst [14, 13] and ourselves [5] for correction of malapropisms and on the ideas of Kondrak [17] for detection of similar words in different languages. The process can be described by the following algorithm:

For each word w in the text being composed
  For each word u looking like w in the source language
    For each translation t of u into the target language
      For each word x in the context of w
        Measure the coherence $C(t,x)$ between t and x
    If $\max_x C(t,x) > \max_x C(w,x)$ then
      Suggest to the user to consider t instead of w

Here, the similarity between strings can be measured as in [17], while the context and coherence measure can be as in [13], [5], or some combination of those.

For example, consider the Spanish phrase *Juan asistió al congreso* 'John attended a conference' and its (bad) translation *John assisted a congress*. Given the latter phrase, the above algorithm will detect a possible correspondence *asistir* (infinitive of *asistió*) to *assist* due to the regular correspondence rules for the two languages. Then, the algorithm will look up in a dictionary the correct translation for the *asistir*, which is *attend*. Then, following [5] and possibly consulting Internet for the statistics, the algorithm will find that *attend the conference* is a better collocation than *assist the conference*. This will allow the algorithm to suggest the translator to substitute *attended* for *assisted* in the target text.

11 Conclusions

Internet provides rich statistics that can help the translator to find which of the possible translation variants is more used in the (foreign) target language, which words can be combined in this language, which prepo-

sitions are used in this language with specific verbs, nouns, or adjectives, as well as to determine or clarify the meaning and/or usage nuances of a foreign words.

However, the use of this enormously rich lexical material is not straightforward. Various precautions are to be taking to avoid false or misleading statistics. The results should always be verified by analyzing a random sample of documents retrieved to check whether the correct word sense, style, or syntactic construction contributes in the final figure.

This complicates fully automatic application of the discussed techniques. However, useful semi-automatic tools can be developed to simplify the routine procedures and fatiguing calculations.

Finally, in addition to the richest lexical statistics, Internet provides the translator with tools and resources that can hardly be found in even a very large library—such as dictionaries and wordlists of different languages and sublanguages, as well as the web interfaces to corpus concordances.

References

1. Apresyan, Yu. D., et al. (Ed.). New Comprehensive English-Russian Dictionary. Russky Yazyk, 1003. CD-ROM edition: English-Russian Electronic Dictionary Multilex 2.0, MediaLingua JSC, 1996, 1997; see www.multilex.ru/online.htm.
2. Banerjeet, S., and T. Pedersen. *The Design, Implementation and Use of the Ngram Statistics Package*. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 4th Intern. Conf. on Computational Linguistics CICLing-2003. Lecture Notes in Computer Science, No. 2588, Springer-Verlag, 2003, p. 372–383.
3. Bolshakov, I. A., A. Gelbukh. *A Large Database of Collocations and Semantic References: Interlingual Applications*. In: M. S. Blekhman (ed.). *Machine Translation. Theory and Practice*. Bahri Publications. New Delhi, 2001, p. 167-187. Also available in: International Journal of Translation, Vol.13, No.1-2, 2001, pp. 167–187.
4. Bolshakov, I. A., A. Gelbukh. *A Very Large Database of Collocations and Semantic Links.* In: Mokrane Bouzeghoub *et al*. (eds.)

Natural Language Processing and Information Systems. Proc. 5th International Conference on Natural Language Applications to Information Systems, NLDB-2000. Lecture Notes in Computer Science No. 1959, Springer-Verlag, 2001, p. 103–114.

5. Bolshakov, I. A., A. Gelbukh. *On Detection of Malapropisms by Multistage Collocation Testing*. Natural Language Processing and Information Systems. Proc. 8th International Conference on Natural Language Applications to Information Systems, NLDB-2003. Lecture Notes in Computer Science, Springer-Verlag, 2003, to appear.
6. Bolshakov, I. A., Sofia N. Galicia-Haro. *Can We Correctly Estimate the Total Number of Pages in Google for a Specific Language?* In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 4th Intern. Conf. on Computational Linguistics CICLing-2003. Lecture Notes in Computer Science, No. 2588, Springer-Verlag, 2003, p. 415–419.
7. Bolshakov, I.A., A. Gelbukh. *Heuristics-based replenishment of collocation databases*. In: E. Ranchhold, N. J. Mamede (Eds.) Advances in Natural Language. Proc. PorTAL-2002: Portugal for Natural Language Processing. Lecture Notes in Computer Science, No. 2389, Springer-Verlag, 2002, p. 25–32.
8. Brill, Eric. *Processing Natural Language without Natural Language Processing*. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 4th Intern. Conf. on Computational Linguistics CICLing-2003. Lecture Notes in Computer Science, No. 2588, Springer-Verlag, 2003, p. 362–371.
9. Chilango *glossary*; see www.ytumamatambien.com/ENGLISH/WEB/glosario.html.
10. Collins Cobuild Corpus Concordance Sampler; see titania.cobuild.collins.co.uk/form.html#queries.
11. Concordance. Text analysis and concordancing software; see www.rjcw.freeserve.co.uk.
12. Gelbukh, A., G. Sidorov, and L. Chanona-Hernández. *Compilation of a Spanish representative corpus*. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 3rd Intern. Conf. on Computational Linguistics CICLing-2002. Lecture Notes in Computer Science, No. 2276, Springer-Verlag, 2002, p. 285–288.

13. Hirst, G., A. Budanitsky. *Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion.* J. Computational Linguistics (to be published).
14. Hirst, G., D. St-Onge. *Lexical Chains as Representation of Context for Detection and Corrections of Malapropisms*. In: C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. The MIT Press, 1998, p. 305–332.
15. *How Do You Say "Hippo" in...*; see www.hippos.com/foreign_ language.htm.
16. Kilgarriff, Adam. *Web as corpus*. Unpublished presentation at the 4th Intern. Conf. on Computational Linguistics CICLing-2003, Mexico City, February 2003; see www.CICLing.org/2003.
17. Kondrak, G. *Identifying Complex Sound Correspondences in Bilingual Wordlists*. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 4th Intern. Conf. on Computational Linguistics CICLing-2003. Lecture Notes in Computer Science, No. 2588, Springer-Verlag, 2003, p. 432–443.
18. *Language Teacher* ER-200D. *Russian-English electronic dictionary and organizer*.
19. Mel'èuk, Igor. *Dependency Syntax: Theory and Practice*. SONY Press, NY, 1988.
20. *Spanish Master* DBE-440. Franklin Bookman electronic dictionary; see www.franklin.com.
21. *The American Heritage Dictionary of the English Language*. Fourth Edition, 2000; see www.bartleby.com/61/46/B0164600. html.
22. *Yandex Search Engine*; see www.yandex.ru.