Identifying the Translations of Idiomatic Expressions using TransSearch

Stéphane Huet, Philippe Langlais

Identifying the Translations of Idiomatic Expressions using TransSearch

Stéphane Huet[1] and Philippe Langlais[2]

[1] LIA - Université d'Avignon, Avignon, France
stephane.huet@univ-avignon.fr
[2] DIRO - Université de Montréal, Montréal, Québec, Canada
felipe@iro.umontreal.ca

Abstract. This document presents a case study relating how a user of TransSearch, a translation spotter as well as a bilingual concordancer available over the Web, can use the tool for finding translations of idiomatic expressions. We show that with some care on the queries made to the system, TransSearch can identify a fair number of idiomatic expressions and their translations.

1 Introduction

Idioms are expressions of a given language, whose sense is not predictable from the meanings and arrangement of their elements [8]. For example, an expression like "*to be hand in glove*" meaning "*to have an extremely close relationship*" cannot easily been deduced from what a hand and a glove are. Some expressions are more analyzable than others; for instance, the meaning of the expression "*fights like cat and dog*" might easily be inferred by the senses of "*cat*" and "*dog*". This is not so for the expression "*it rains cats and dogs*". In this work, we are interested in identifying the translation of this second type of expressions.

Idioms — and more generally Multi-Word Expressions (MWEs) — pose significant problems for many applications of natural language processing since they are numerous in most languages and have idiosyncratic meanings that severely disturb deep analysis [11]. The problem of MWEs — and idioms in particular — is especially acute in the case of Machine Translation (MT) where a failure of the system to detect such expressions often leads to unnatural, if not hilarious outputs.

Therefore, one important component of an MT system is its lexicon of MWEs. This is true for rule-based MT systems as well as statistical MT (SMT) ones. Currently, state-of-the-art phrase-based SMT systems rely on models (pairs of phrases) that do not handle MWE specifically. Some authors have been trying to group multi-word expressions before the alignment process [4] or to add a new feature encoding the knowledge that a given phrase pair is a MWE [10, 2]. This last work showed that using manually defined WordNet MWEs could improve MT.

Not only are idioms interesting for improving MT systems, they are as well notably known to pose problems to non-native persons. This is especially true

when a second-language idiom is much different from its translation into the native language. For instance, French speakers might easily catch the English idiom "*play cat and mouse*" because its French translation "*jouer au chat et à la souris*" is literal in this case. On the contrary, they could find hard to understand "*He couldn't say boo to a goose*"[3] because its translation into French "*Il est d'une timidité maladive*" (literaly "*He is sickly shy*") is completely different.

Idiomatic expressions are interesting for professional translators as well. In [6], the authors analyzed the most frequent queries submitted by users to the bilingual concordancer TransSearch. They found that among others things, users frequently queried idiomatic phrasal verb expressions, such as "*looking forward to*". Because they were expecting that the users would query idiomatic expressions, they did not investigate further this aspect of the logfile, but instead concentrated on analyzing the prepositional phrases (some of which being idiomatic) frequently submitted to the system.

In this paper, we study the problem of translating idiomatic expressions from a user perspective. We tried to identify the translations of a number of idioms in the Translation Memory (TM) of the new version of the bilingual concordancer TransSearch. Since many idioms have inflected forms, we show the impact of different strategies for querying the database. For instance, in the (idiomatic) expression "*to keep oneself to oneself*", both the verb "*keep*" and the pronoun "*oneself*" can vary according to conjugation and inflection respectively, and verbatim queries may fail to identify relevant occurrences of the expression.

The remainder of the paper is organized as follows. Section 2 describes TransSearch, the Web application we employed in our experiments. Section 3 presents the data we used and how we submit queries to the TM system to find translations. Section 4 is dedicated to the evaluation of the translations proposed by the system, while Section 5 concludes and explores further perspectives.

2 TransSearch

TransSearch is a bilingual concordancer that allows its users to query large databases of past translations in order to find ready-made solutions to a host of translation problems. Subscribers of the system are mainly professional translators. A recent study of their query logs exhibits that TransSearch is used to answer difficult translation problems [6]. Among the 7.2 million queries submitted to the system over a six-year period, 87% contain at least two words. Among the most frequent submitted queries, several appear to be idiomatic, like "*out of the blue*" or "*in light of*".

2.1 System Features

Made available since 1996 through a Web interface by the Université de Montréal [7], TransSearch has recently been improved to become not only a bilingual concordancer but also a translation finder [1]. Figure 1 which displays the

[3] At the time of writing, *Google Translate* produces the literal translation "*Il ne pouvait pas dire boo à une oie*".

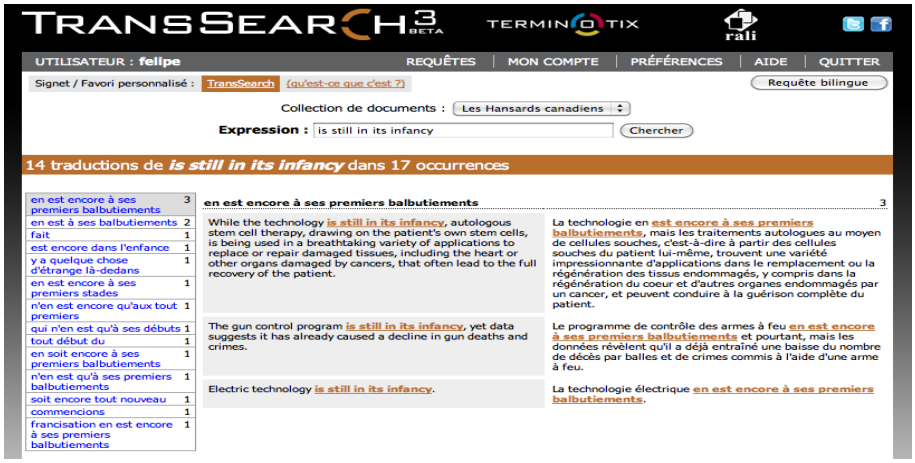

Fig. 1. Result returned by the new TransSearch to the query "*is still in its infancy*". The left column shows likely translations in decreasing order of likelihood, while the main columns shows concordances. The query and the selected translation are shown in color in each of them.

results for the query "*is still in its infancy*" exemplifies the new capabilities of the system. Where a simple bilingual concordancer (as were the previous versions of TransSearch) would only display a list of parallel sentences containing the query in their English part, the new version of TransSearch highlights for each sentence pair the French part associated with the query. Besides, this version displays on the left hand side the whole range of translations (automatically) found in the TM. For the first suggested translation, "*en est encore à ses premiers balbutiements*", three of the sentence pairs containing a variant of this translation (see the merging process described in Section 2.2) are displayed in context.

With respect to an ordinary bilingual concordancer, where the identification of translations in sentences is left to the user, we believe the new version of TransSearch dramatically improves usability, by displaying a general view of the TM content for a given query.

The previous query example has shown that the system is able to find results for queries with several words. The user can also submit more advanced queries to search discontinuous expressions. For example, Figure 2 displays the results for the query "*make .. hair stand on end*". The '..' operator enables the user to indicate the system that occurrences of 2 words in the query (here "*make*" and "*hair*") can be up to 5 words apart inside a sentence. Another operator '...' allows for searches without constraining the distance between two words. From a linguistic perspective, these two operators are useful since they enable the user to spot expressions where words may be separated by a few words, such as nominal groups in the examples of Figure 2.

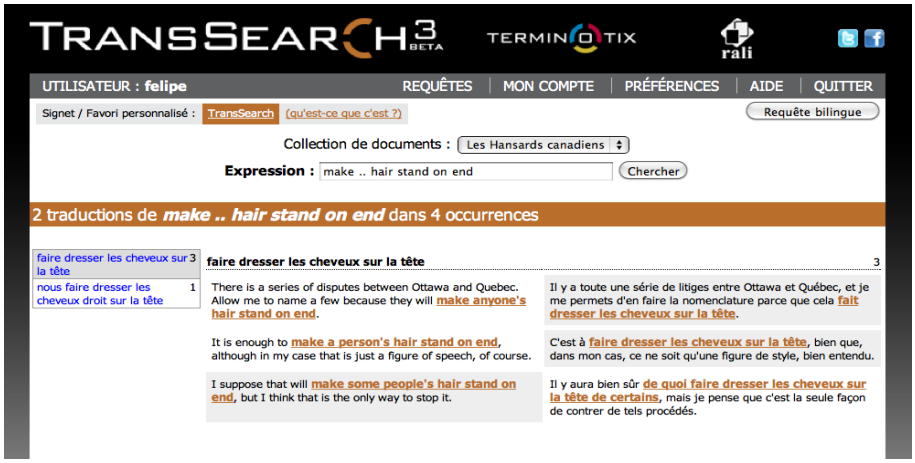

Fig. 2. Result returned by TransSearch to the query "*make .. hair stand on end*".

Besides, another advanced type of query is available in TransSearch: morphological expansions. The system considers all the morphological derivations of the terms associated with the '+' symbol, when retrieving sentence pairs. Figure 3 shows the results for the query "*take+ no for an answer*". In this example, the interface displays expressions containing different inflected forms of the verb "*take*". This last operator is specially useful for morphologically rich languages like French or Spanish and allows the user to spot translations without taking care of their possible inflections.

By default, TransSearch searches for the given expression regardless of languages (French or English). In some cases however, it is necessary to specify the language, for instance in order to distinguish between the French and English words "*tape*" ("*to hit*" in French). Using the same mechanism, it is also possible to look up occurrences of a specific translation of a given query by filling at the same time the French and English fields of the query form. For example, a user can check that "*les dés sont pipés*" is a correct translation of "*the dice are loaded*" by looking at the same time at these two expressions into the TM sentence pairs.

2.2 Processing Steps

In order to suggest several translations for a given query, TransSearch performs several processing steps that we briefly describe hereafter. Many current computer-assisted translation tools mainly rely on sentence-level matching to exploit their translation memory. TransSearch operates at a finer-grained level using word alignment techniques, which are commonly used in SMT. The term translation spotting, coined by [13] and relabeled here as *transpotting*, is defined as the task of identifying the target language word-tokens that correspond

Fig. 3. Result returned by TRANSSEARCH to the query "*take+ no for an answer*".

to a given source language query in a pair of sentences known to be mutual translations; it is a core step in the new version of TRANSSEARCH.

We call *transpot* the target word-tokens automatically associated with a query in a given pair of sentences. For instance in Figure 1, "*en est encore à ses premiers balbutiements*" and "*soit encore tout nouveau*" are 2 out of 14 distinct transpots displayed to the user for the query "*is still in its infancy*".

The method used to transpot queries in the retrieved sentence pairs is described in details elsewhere. In a nutshell, our transpotting algorithm uses statistical word-alignment models and enforces that the transpots identified are sequences of contiguous words. As mentioned in [12], contiguous tokens in the source language sentence tend to be aligned to contiguous tokens in the target language. This statement is confirmed by the good experimental results presented in the study of [1].

Queries that occur frequently in the TM receive numerous translations using the transpotting methods described above, some being clearly wrong, some others being redundant (morphological variations of the same translation). We estimate that a user will focus on the 10 first translations presented, so we want to provide as many correct and diversified translations as possible at the top of the result page. Therefore, two postprocessing steps were introduced inside the TRANSSEARCH engine. The first one filters out bad transpots using supervised learning. A classifier was trained on a corpus where transpots were manually labeled as "good" or "bad", using features such as the ratio of grammatical words inside the hypothesized transpots. Once transpots have been filtered out, the second step merges those which are different inflectional forms of the same sequence of canonical words. For instance, "*au nom du*" and "*au nom des*" will be considered as similar, since "*du*" and "*des*" are contractions of "*de + le*" and "*de + les*" respectively, where "*le*" and "*les*" are definite articles. Furthermore,

as it was noticed that translations that differ only by a few grammatical words or punctuation marks, like "*de la part de*" and "*part de*" are often redundant for the user, those are combined as well. At the end of this second post-processing step, only the most frequent transpot of each merged set is displayed on the left hand side of the user interface (see Fig. 1 to 3). These transpots are shown as a list sorted in the decreasing order of their transpotting frequency.

3 Methodology

3.1 Resources

Translation Memory The largest TM used in TransSearch comes from the Canadian Hansards, a collection of the official proceedings of the Canadian Parliament. For our experiments, we used an in-house sentence aligner [5] to align 8.3 million French-English sentence pairs extracted from the 1986-2007 period of the Hansards. This bitext was indexed with Lucene[4] to form our TM.

Idiom Lexicon Classifying an expression as idiomatic or not is not an easy task. Therefore, we resorted to the phrase book [9] written by Jean-Bernard Piat, a translation teacher as well as a translator. This book oriented towards general public market provides a list of 1,467 idiomatic expressions in both languages (French and English) categorized by subjects (e.g. human body).

According to the author, the expressions were chosen because they are frequently used. A minority of these expressions are expressed in an informal language (e.g. "*to be well-upholstered*"). He also mentioned that it happens sometimes that an idiomatic expression in one language (e.g. "*to burn the midnight oil*") is not idiomatic in the other language (e.g. "*travailler tard dans la nuit*").

Examples of entries in this book are reported in Table 1. A few entries have several equivalent translations such as "*make your flesh creep*" and "*give you goose pimples*" for "*donner la chaire de poule*". Globally, there are on average 1.17 English translations and 1.01 French translations per entry.

All expressions but seven, are used in the context of a sentence. According to the author, using expressions in a context makes them easier to understand and to use for the readers. The lexicon contains a high proportion of verbal phrases (around four out of five of the available entries) that are used in their inflected form, like "*He took to his heels*" for the phrase "*to take one's heels*". Other entries are fixed expressions such as "*When there's a will, there's a way*" or "*Hands off!*".

3.2 Preprocessing

In order to take into account contextualization that makes lexicon entries too specific, the used lexicon was manually annotated by the first author of this paper.

[4] http://lucene.apache.org

Table 1. Excerpt of the entries we considered in our experiment. R stands for the reference translation, G stands for the translation made by Google Translate (provided as a proxy to literal translation). Words in parenthesis have been manually marked as contextual words that are not part of the idiomatic expression.

French	English
Il est agile comme un singe	R *He's as nimble as a goat* G *He is agile as a monkey*
Elle était sur son trente et un	R *She was dressed to kill* R *She was all dressed up* G *She was on her thirty-one*
(Je vais d'abord) me rincer la dalle — familiar —	R *(I'm going to) wet my whistle (first)* G *First I'll rinse my slab*
(Il aime) rouler des mécaniques — familiar —	R *(He likes) flexing his muscles* R *(He likes) playing the tough guy* G He loves rolling mechanical
J'ai vu trente-six chandelles	R *I saw stars* G *I saw thirty-six candles*

All words judged as extra-information with respect to the idiomatic expression were annotated as such in the lexicon. Those are the words in parenthesis in the examples of Table 1. They are typically modal verbs (e.g. "*can*", "*must*"), semi-modal verbs (e.g. "*am going to*", "*are likely to*"), catenative verbs (e.g. "*want to*", "*keep*"), adverbs (e.g. "*only*", "*finally*"), adverbial phrases (e.g. "*in Italy*", "*when he heard the news*") or noun phrases (e.g. "*this poet*", "*his latest book*"). Finally, at least one word was classified as extra-information for 486 out of 1,467 entries.

3.3 Queries to the Translation Memory

In order to test the ability of TransSearch to find translations for idioms, three types of queries were submitted to the system: queries built from either the English side or the French side of the entry, and bilingual queries where both sides where searched for at the same time. As mentioned in Section 3.1, a few entries have more than one English or French reference translations. For these entries, we collected results found from all the equivalent translations. Since TransSearch user interface doest not allow users to write an "or" operator between several equivalent translations, we had to simulate the behavior of this operator by submitting independently translations and then by merging results retrieved by TransSearch.

Table 2 shows the number of lexicon entries found in the TM, using bilingual (column 2), English (column 3) or French queries (column 4) and considering various ways of querying the system. As expected, building verbatim queries from the lexicon leads to retrieve information inside the TM for a small number

Table 2. Number of the lexicon entries found inside the translation memory using several types of query.

Query types	bilingual	English	French
verbatim queries	37	136	248
EN: *I have no axe to grind*			
FR: *Je ne prêche pas pour ma paroisse*			
+ manual removal of extra words	91	302	410
EN: *I have .. axe to grind*			
FR: *Je .. prêche .. pour ma paroisse*			
+ removal of extra pronouns	106	381	509
EN: *have .. axe to grind*			
FR: *prêche .. pour ma paroisse*			
+ verb lemmatization	210	624	650
EN: *have+ .. axe to grind*			
FR: *prêcher+ .. pour ma paroisse*			
+ pronoun and determiner lemmatization	238	700	705
EN: *have+ .. axe to grind*			
FR: *prêcher+ .. pour sa+ paroisse*			

of expressions only (line 1). After taking into account the manual preprocessing step introduced in Section 3.2, that is, after removing extra words, nearly three times as many queries have at least one hit in the TM (line 2). Still, at best, a user could retrieve no more than 410 (French) expressions from the 1,467 ones by simply querying them verbatim or removing extra words.

An inspection of the submitted queries revealed that many of them correspond to flexible idioms, that is, idiomatic expressions that can vary from one occurrence to another. In order to capture those variations and to increase therefore the number of hits in the TM, we wrote rules that abstract away some of those variations. For this, we used a mix of linguistic information as well as the operators we described earlier. We resisted to the temptation of adjusting this process for each query and instead applied some rules in a systematic way, given a set of linguistic markers semi-automatically annotated in the lexicon.

The performed processing steps for the entry [“*I have no axe to grind*”, “*Je ne prêche pas pour ma paroisse*”] are illustrated in Table 2. A set of rules deleted personal pronouns at the beginning of an expression (see line 3); a list of pronouns to be removed has been collected for this in each language. Then, lemmatized verbs were replaced by the corresponding lemma and auxiliary verbs were removed (see line 4); we used for this an in-house lemmatization resource available for both languages. Last, we also considered lemmatizing pronouns and determiners within an expression (see line 5).

It should be noted that we chose to modify entries using a set of limited rules in order to avoid over-abstracting idiomatic expressions. For instance, we noticed that the indefinite pronoun "*it*" in English usually occurs in fixed expressions and cannot be replaced by another personal pronoun. Therefore, we kept this pronoun verbatim in the queries made.

We observe in Table 2 the dramatic increase of the number of hits in the TM according to the level of abstraction of the query. At best, the rewriting rules we applied allow TransSearch to return sentence pairs for 700 English entries and for 705 French entries, i.e. roughly half of the lexicon. Each set of rules increases the number of queries with at least one hit. Surprisingly, verb lemmatization led to a higher improvement of the coverage for English queries than for French ones. This shows that, on the contrary to what we expected first, this process is also relevant for weakly inflected languages.

This experiment also shows that in order to get the best of the system, users should use the linguistic operators at their disposal. Since we know that most queries made by real users of the application do not use those operators it could mean one of two things. When users submit a query to the system without getting any answer, they might simply abandon the search for a translation or on the contrary, they might figure out a way to process the query in order to find a match in the TM. Inspecting the log-files of the application exhibits evidences that both strategies happen in practice. This means that automatically processing the query of a user is an interesting prospect to consider.

Another interesting outcome of the experiment we conducted is that the Hansards indexed by TransSearch are rather good for identifying the idiomatic expressions we considered.

4 Evaluation

We have measured the quantity of idiomatic expressions we could find by querying the Hansards indexed by TransSearch. We now turn to the evaluation of how good the application is for spotting the translations of the retrieved expressions. Once again, it should be noted that in most bilingual concordancers we know of, this part is left to the user.

4.1 Objective Evaluation

For the French and English queries obtained after applying our rewriting rules, TransSearch was able to retrieve on average respectively 36.1 and 31.7 sentence pairs from the TM. Among this material, the transpoting algorithm identified respectively 12.5 French and 14.9 English (different) translations (shown to the user on the left of the navigator). Since a manual analysis of all the suggested translations would be a tedious task, an evaluation was performed thanks to the sanctioned translations belonging to the idiom lexicon described in Section 3. As shown in Table 2 (last line), a query and its sanctioned translation are found simultaneously in the sentence pairs returned by the system for 238 lexicon entries.

Table 3. Recall (%) measured using the lexicon sanctioned by the translation memory as a reference.

k	1	2	3	5	10	all
English queries	41.6	56.3	59.2	65.1	69.3	74.8
French queries	41.6	49.6	54.6	62.6	69.3	76.5

Therefore we restrained our objective evaluation to those 238 queries. Table 3 provides the proportion of those queries where the k-first translations displayed by TransSearch contain (at least) one of the reference translations sanctioned by the lexicon.[5]

The recall of 75% measured when all the translations returned by the system are considered demonstrates that the embedded transpotting algorithm has the ability to find translations in the retrieved sentence pairs. The result of 41,6% obtained when considering the first translation returned by the system (that is, the most frequent one) is not bad either, especially since the reference we used is rather incomplete. For instance, our lexicon contains the translation "*être dans un état second*" for the idiom "*to be in a daze*", while TransSearch displays this translation after "*est nébuleux*", which is as well a good translation of the English idiom. Similarly, TransSearch returns no less than 34 different translations[6] of the query "*be+ around the corner*", most of which being perfectly legitimate translations, while our reference contains only one.

4.2 Manual Evaluation

The objective evaluation revealed the great potential of TransSearch for identifying the translation of idiomatic expressions, but also showed that a manual evaluation was required in order to account for the sparseness of our bilingual lexicon. Therefore, we conducted a manual evaluation involving 5 bilingual annotators that were presented with lists of identified translations among 100 randomly chosen French queries. They were asked to indicate in those lists those translations that they found correct, partially correct or wrong. No specific guidelines were given to explain these labels. At the time of writing, 50 queries were judged by 3 annotators and the 50 other by 2.

Globally, the quality appreciated by the annotators turned out to be variable, some annotators tending to classify more easily translations as correct. This translated into a low value of 0.25 obtained when computing the Fleiss inter-annotator agreement [3]. Figure 4 illustrates some cases of divergence.

The results of this evaluation are reported in Table 4. Since a given query can be rated differently by several judges, we credited divergent annotations equally.

[5] In order to account for inflectional variations, we compared lemmatized translations.

[6] The 10 most frequent ones are: *est à nos portes*, *arrive à grand pas*, *était imminent*, *nous attend*, *me guette*, *est sur le point*, *s'annonce*, *est en vue*, *sommes au bord de*, and *survenir*.

appeler un chat un chat	J1	J2	J5
▷ we should call it what it is	correct	correct	correct
▷ we can say the d word and the m word	correct	wrong	partial
▷ calling manure a rose doesn't change the smell	correct	wrong	partial

manger à tous les râteliers	J1	J2	J5
▷ slurps at everyone 's trough	correct	correct	correct
▷ double - dipper	partial	correct	partial
▷ them pot lickers and accusing them of being at the trough and pork barrelling	wrong	partial	wrong

Fig. 4. Examples of annotations of some French idiomatic queries.

For instance, if a translation is judged correct by one annotator, and wrong by another one, a credit of 0.5 will be given to each label.

For all but 7 queries, TransSearch is able to identify a translation classified as correct by at least one annotator. For these queries, the average rank of the first correct translation is 1.4, which indicates that relevant translations can usually be found among the two first displayed by TransSearch. Also, on average, we observe that only 36% of the translations proposed to the user are labeled as wrong.

Table 4. Average percentage of translations judged correct, partially correct or wrong per query on a sample of 100 English queries randomly selected. *avr* stands for the average number of translations produced per query, while *rank* indicates the average rank of the first translation labeled as correct by at least one annotator.

correct	partial	wrong	*avr*	*rank*
42%	22%	36%	13.4	1.4

5 Conclusion

In this work, we have studied the problem of identifying translations of idiomatic expressions in both English and French, using a brand new version of the bilingual concordancer TransSearch. We showed that a user that would query the system verbatim would often fail to find a match in the TM and that some cleverness is required in order to get good use of the system, such as resorting to the morphological (+) and the proximity (..) operators available in the query language recognized by the system. We automatized the querying process and showed that a rough half of the idiomatic expressions queried to the system finally got a match in the TM, while a high proportion of the translations returned by the system are correct.

Acknowledgments

This work was funded by an NSERC grant in collaboration with Terminotix.[7] We are indebted to Sandy Dincky, Fabienne Venant and Neil Stewart who kindly participated to the annotation task.

References

1. Julien Bourdaillet, Stéphane Huet, Philippe Langlais, and Guy Lapalme. TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation Journal*, 24(3–4):241–271, 2010.
2. Marine Carpuat and Mona Diab. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proceedings of NAACL-HLT*, pages 242–245, Los Angeles, CA, USA, 2010.
3. Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Pai. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.
4. Patrik Lambert and Rafael Banchs. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of MT Summit*, pages 396–403, Phuket, Thailand, 2005.
5. Philippe Langlais. A system to align complex bilingual corpora. Technical report, CTT, KTH, Stockholm, Sweden, 1997.
6. Elliott Macklovitch, Guy Lapalme, and Fabrizio Gotti. TransSearch: What are translators looking for? In *Proceedings of AMTA*, pages 412–419, Waikiki, Hawaii, USA, 2008.
7. Elliott Macklovitch, Michel Simard, and Philippe Langlais. TransSearch: A free translation memory on the World Wide Web. In *Proceedings of LREC*, pages 1201–1208, Athens, Greece, 2000.
8. Tom McArthur, editor. *The Oxford Companion to the English Language*. Oxford University Press, 1992.
9. Jean-Bernard Piat. *It's raining cats and dogs et autres expressions idiomatiques anglaises*. Librio. J'ai lu, 2008.
10. Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL-IJCNLP Workshop on Multiword Expressions*, pages 47–54, Suntec, Singapore, 2009.
11. Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15, Mexico City, Mexico, 2002. Springer.
12. Michel Simard. Translation spotting for translation memories. In *Proceedings of the HLT-NAACL Workshop on Building and using parallel texts: data driven machine translation and beyond*, volume 3, pages 65–72, Edmonton, Canada, 2003.
13. Jean Véronis and Philippe Langlais. *Evaluation of Parallel Text Alignment Systems — The Arcade Project.*, chapter 19, pages 369–388. Kluwer Academic Publisher, 2000.

[7] `www.terminotix.com`