Translation of Idiomatic Expressions Across Different Languages: A Study of the Effectiveness of TransSearch

Stéphane Huet, Philippe Langlais

Translation of Idiomatic Expressions across Different Languages: A Study of the Effectiveness of TRANSSEARCH

Stéphane Huet and Philippe Langlais

Abstract This chapter 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 by paying close attention to the queries made to the system, TRANSSEARCH can effectively identify a fair number of idiomatic expressions and their translations. For indicative purposes, we compare the translations identified by our application to those returned by GOOGLE TRANSLATE and conduct a survey of recent Computer-Assisted Translation tools with similar functionalities to TRANSSEARCH.

1 Introduction

Idioms are commonly defined as expressions of a given language, whose sense is not predictable from the meanings and arrangement of their elements [13]. For example, an expression like "*to be hand in glove*" meaning "*to have an extremely close relationship*" cannot have easily been deduced from what a hand and a glove are. 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 [20]. 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 comical 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)

Stéphane Huet
LIA-CERI — Université d'Avignon, Avignon, France, e-mail: stephane.huet@univ-avignon.fr

Philippe Langlais
DIRO — Université de Montréal, Montréal, Québec, Canada, e-mail: felipe@iro.umontreal.ca

that do not handle MWE specifically. Some authors have been trying to group multi-word expressions before the alignment process [9] or to add a new feature encoding the knowledge that a given phrase pair is a MWE [19, 5]. These two last works showed that MT could be improved with MWEs extracted automatically, or defined manually from WORDNET.

Not only are idioms interesting for improving MT systems, they are well known to pose problems to non-native speakers. 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*"[1] because its translation into French "*Il est d'une timidité maladive*" (literally "*He is sickly shy*") is completely different.

Idiomatic expressions are interesting for professional translators as well. In [11], 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 this aspect of the logfile any further, but concentrated instead on analyzing the prepositional phrases (some of which were idiomatic) frequently submitted to the system.

In this paper, we study the problem of translating idiomatic expressions from a user perspective. We attempted 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 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 presents the variability of idiomatic expressions and the interest of Computer-Assisted Translation (CAT) tools for users to translate them. Section 3 describes TRANSSEARCH, the Web application we employed in our experiments. Section 4 provides information about the data we used and the query submission process to the TM system to find translations. Section 5 is dedicated to the evaluation of the translations proposed by the system, including the comparison of TRANSSEARCH with GOOGLE TRANSLATE. Section 6 conducts a survey of recent CAT tools with similar functionalities to TRANSSEARCH. Section 7 provides a conclusion.

[1] At the time of writing, GOOGLE TRANSLATE produces the literal translation "*Il ne pouvait pas dire boo à une oie*".

2 Idiomatic expressions and CAT tools

2.1 What is an idiomatic expression?

It is difficult to find a universal definition that covers the variety of what can constitute idiomatic expressions, examples of which are "*give up*" and "*his ears must be burning*". As mentioned at the beginning of this chapter, they are often defined as sequences of words involving some degrees of semantic idiosyncrasy or non-compositionality.

In phraseology, idiomatic expressions — also named phraseological expressions or phrasemes — are defined as non-free multi-word expressions, which means that at least one of their components is selectively constrained or restricted by linguistic convention such that it is not chosen freely [14, 15]. For example, the expression "*be in the same boat*" , meaning "*have the same problem*", is syntactically and morphologically organized as any English phrase and can even mean "*be on a boat*". What makes it special is the fact that it has an unpredictable sense and has components which cannot be replaced by any synonym (e.g. "*boat*" by "*ship*") without removing the distinctive idiomatic meaning.

In [15], Mel'čuk separates phrasemes into two types of expressions based on whether they are defined at the pragmatic or semantic level. Pragmatic phrasemes (or pragmatemes) are produced when all the components of the expression are constrained by the situation. For example, a sign that informs car drivers they may not park in a given place should use the idiomatic expression "*No parking*" rather than the non-idiomatic "*Parking forbidden*" [18].

Semantic phrasemes, on the other hand, are produced when the choice of a meaning from a given conceptual representation is free but the selection of at least one component of the expression is not free. Semantic phrasemes include three main categories: clichés, collocations and idioms.[2]

Clichés and collocations are compositional, i.e. for a given semantic phraseme AB, the meaning and the form of A and B are combined in accordance with the rules of the language. On one hand, clichés are phrasemes where none of the components is selected freely, i.e. cannot be replaced by a (quasi-)equivalent expression. For example, "*something*" cannot be used instead of "*one thing*" in "*one thing after another*", while "*we all produce mistakes*" or "*we all make errors*" can be understood but are not as natural as "*we all make mistakes*". On the other hand, collocations have one component[3] chosen freely by the speaker and another component chosen as a function of the base. To characterize for example a battle as being very violent, "*fierce* BATTLE" is more standard than "*ferocious* BATTLE" or "*terrible* BATTLE", while "*award a* PRIZE *to*" will be used to express "*give a* PRIZE *to*".

Unlike clichés and collocations, idioms are non-compositional and none of their components is selected freely. In this work, we are interested in identifying the trans-

[2] Idioms — named *locutions* in French — are seen in phraseology as a subcategory of phrasemes and are used in the remainder of this paper as a synonymous of idiomatic expressions.

[3] Shown in small caps in the examples.

lation of this last category: idioms. Idioms can differ on the degree of transparency, the degree to which their meaning includes the meanings of their components. Here are some examples: "*let's go Dutch*", "*as well as*" or "*throw up*" each has a meaning which does not include the meaning of one of its components ("*go*" and "*Dutch*" in the first example); "*heavy water*" or "*sea anemone*" include the meaning of only one of their lexical components (here "*water*" and "*sea*") but not as their semantic pivot; "*start a family*" or "*shopping center*" include the meaning of all their components but have an additional unpredictable meaning ("*start a family*", for example, means that a new family comes to existence but also that a first child was conceived with one's spouse). These various degrees of semantic analyzability and semantic decomposability make idioms difficult to be identified by automatic methods or even by human annotators [6].

2.2 Finding translations of idioms

For a human wishing to translate an idiom, probably the most natural way is to look it up in a dictionary. This may be difficult because idioms are so numerous and they are not all covered in a given dictionary. Mel'čuk, for example, estimated that the number of non-compositional idiomatic expressions is between 10,000 and 20,000 for any given language [15], while other idiomatic expressions like collocations suffer from a lexical proliferation problem (e.g. "*take a walk*", "*take a hike*", "*take a trip*"...) and are much more numerous [20].

An alternative resource is a translation memory. TMs are databases that store sentences pairwise from the source and target languages. They are typically made of sentences previously translated by professional translators, which makes them more reliable than MT systems. Thus, they represent a valuable resource for translating idioms especially when they store a huge quantity of parallel corpora [24]. Many commercial CAT tools, such as SDL TRADOS[4], DEJA VU[5], LOGITERM[6] or MULTITRANS[7], are available to manage and search information in a TM. They mainly operate at the level of sentences, which limits their usefulness to repetitive translation tasks. As we shall see in Section 6, not all TMs have this limitation. Tools such as TRANSSEARCH or TRADOOIT are able to operate at the word level since they typically embed word-alignment technology.

Searching a fixed idiom (e.g."*of course*" or "*till kingdom come*") is straightforward since it always occurs in the same form. Unfortunately, most idioms, in particular expressions of the type "Verb+Noun" are syntactically well-formed phrases that allow some variability in expression [6]. There are several patterns of variability: many inflection forms can occur for a given idiomatic expression, such as "*have*

[4] http://www.trados.com.

[5] http://www.atril.com.

[6] http://terminotix.com.

[7] http://www.multicorpora.

other fish to fry" whose verb can be conjugated for person ("*have*","*has*"), tense ("*had*","*will have*"...) or mood ("*would have*", "*having*"...). Some can also undergo passivation, or speaking in the passive voice, like "*the breeze was shot*" or topicalisation like "*it is these strings that he pulled*". Making matters even more complicated is that some words can be inserted within the construction of idioms, such as "*exact*" in the expression "*go exact halves*", which is particularly prominent in languages such as German where verbs are often detached from their arguments [1]. Lastly, some idioms in French, for example, can even allow some semantic replacements, such as "*louper*" or "*manquer*" which can be used instead of "*rater*" in the expression "*rater le coche*".

Given that completely fixed idioms can be represented by a sequence of space-separated words in dictionaries and be searched verbatim inside a TM, we depend on more elaborate strategies for flexible expressions. To guide this search, a system may resort to detailed descriptions of known idioms. The PHRASE MANAGER system presented in [24] was for instance designed to identify multiword expressions during dictionary look up, asking users to provide for each new idiom the headphrase in canonical form as well as the morphological restrictions for each individual word, and to assign it to a syntactic class specifying its possible transformations. Another solution previously proposed relies on the use of Part-of-Speech (PoS) taggers and morphological analyzers [22]. This system standardizes words by taking the basic form of verbs (infinitive), personal pronouns ("*one*", "*oneself*") and possessive pronoun ("*one's*"), while the articles are expanded to match an occurrence of "*a*" with an entry having "*the*". It also resorts to rules learned on a small corpus in order to allow insertions of words of some PoS categories, according to the PoS sequence of the idiom.

In the following section, we present the bilingual concordancer TRANSSEARCH, which exhibits interesting properties when identifying the translation of a given query whether idiomatic or not.

3 TRANSSEARCH

TRANSSEARCH is a bilingual concordancer that allows its users to query large databases of past translations in order to find ready-made solutions for a host of translation problems. Subscribers to the system consist mainly of professional translators. A recent study of the query logs of this application exhibits that TRANSSEARCH is used to answer difficult translation problems [11]. Among the 7.2 million queries submitted to the system over a six-year period, 87% contain at least two words. Among the most frequently submitted queries, several appear to be idiomatic, like "*in light of*" (544 times) or "*out of the blue*" (508 times).

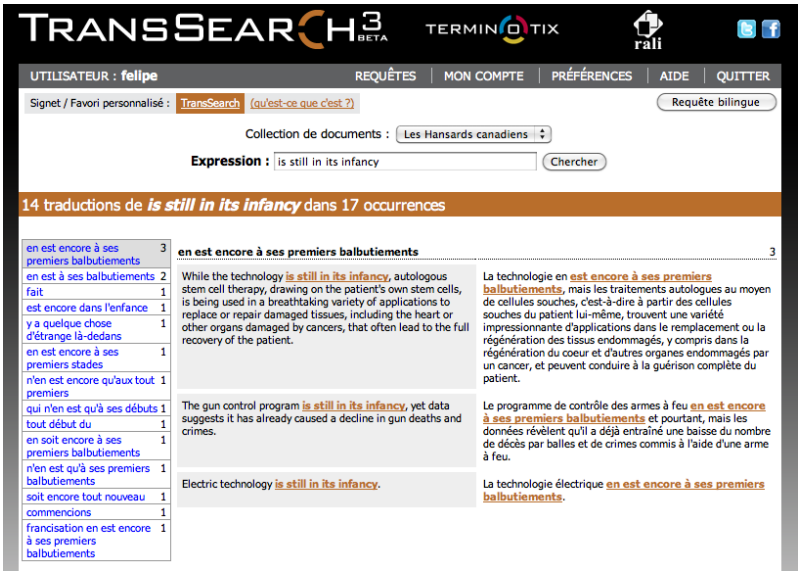

Fig. 1 Result returned by the new TRANSSEARCH to the query "*is still in its infancy*". The left column shows translations from the most likely to the least likely, while the main columns shows concordances. The query and the selected translation are shown in color in each of them. The highlighted translations are hyperlinks to their occurrence in the original Hansards session.

3.1 System Features

TRANSSEARCH, which has been made available since 1996 through a Web interface by the Université de Montréal [12], has evolved into not only a bilingual concordancer but also a translation finder [2]. Figure 1 which displays the results for the query of the idiomatic expression "*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 simply 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 3.2) 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

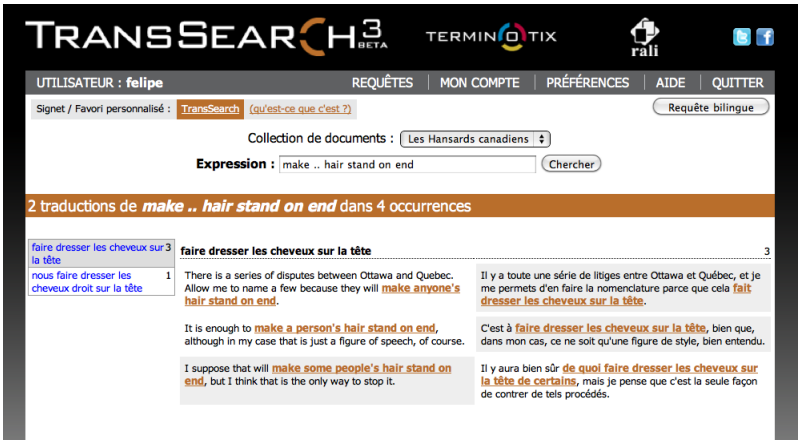

Fig. 2 Result returned by TRANSSEARCH to the query "*make .. hair stand on end*".

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.

Another advanced type of query that is also available in TRANSSEARCH is known as morphological expansions. The system thus 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 especially useful for morphologically rich languages like French or Spanish and allows the user to spot translations without having to worry about all possible inflections for such expressions.

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 both 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 both expressions in the TM sentence pairs.

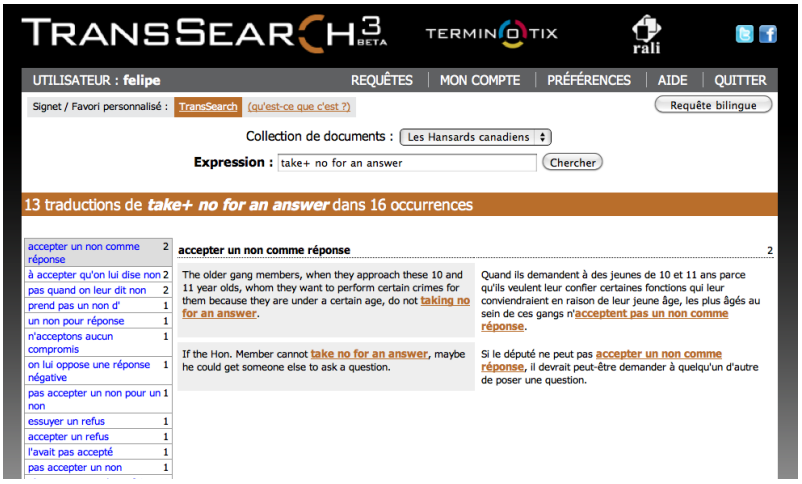

Fig. 3 Result returned by TRANSSEARCH to the query "*take+ no for an answer*".

3.2 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 Véronis and Langlais [23] and relabeled by the authors as *transpotting*, is defined as the task of identifying the target language word-tokens that correspond 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 [2]. In sum, our transpotting algorithm uses statistical word-alignment models and enforces that the transpots identified are sequences of contiguous words. As mentioned in [21], contiguous tokens in the source language sentence tend to be aligned with contiguous tokens in the target language. This statement is confirmed by the good experimental results presented in the study of [2].

Queries that occur frequently in the TM receive numerous translations using the transpotting methods described above, some of course being clearly wrong; others being redundant (morphological variations of the same translation). We estimate that since a user will focus on the 10 first translations presented, 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. To do this, 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 we 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 translations 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.

4 Methodology

The relevance of the results produced by the TRANSSEARCH engine is closely related to the indexed translation memory. This section presents the data used by our system and describes how queries were submitted to test the behavior of the concordancer for idiomatic expressions.

4.1 Resources

4.1.1 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 [10] to align 8.3 million French-English sentence pairs extracted from the 1986-2007 period of the Hansards. This bitext was indexed with Lucene[8] to form our TM. Let us note that this corpus, although it is produced in Canada, can be mainly seen as expressed in a 'standard' French rather than in a Canadian variety of the French language. Sentences are uttered in a formal context and contain very few typical Canadian expressions with respect to everyday French Canadian spoken language.

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

4.1.2 Idiom Lexicon

As mentioned above, determining whether an expression is idiomatic or not is certainly not an easy task. Therefore, we employed the phrase book [17] written by Jean-Bernard Piat, a translation teacher as well as a translator. This book which is oriented towards general public 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 informal (e.g. "*to be well-upholstered*"). He also mentioned that sometimes he could not find an idiom (e.g. "*travailler tard dans la nuit*") in one language to translate idiomatic expressions from the other language (e.g. "*to burn the midnight oil*").

Examples of entries in this book are reported in Table 1. In order to illustrate the problem with translating those expressions, we provide the translations produced by GOOGLE TRANSLATE. A few entries have several equivalent translations such as "*make your flesh creep*" and "*give you goose pimples*" for "*donner la chair 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, providing expressions in context makes them easier to understand and to use. 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!*".

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 (which provides here 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	R	*(I'm going to) wet my whistle (first)*
— familiar —	G	*First I'll rinse my slab*
(Il aime) rouler des mécaniques	R	*(He likes) flexing his muscles*
— familiar —	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*

4.2 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. 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.

4.3 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 were searched at the same time. As mentioned in Section 4.1, a few entries have more than one English or French reference translations. For these entries, we collected results from all the equivalent translations. Since the TRANSSEARCH user interface does not allow users to write an "or" operator between several equivalent translations, we had to simulate the behavior of this operator by submitting independent translations and then by merging the 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 of expressions only (line 1). After taking into account the manual preprocessing step introduced in Section 4.2, that is, after removing extra words, twice as many queries had at least one hit in the TM (line 2). Still, at best, a user could retrieve no more than 28 % of the French expressions by simply querying them verbatim or by 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 the number of hits in the TM, we used a mix of linguistic information as well as the operators we described earlier. In so doing, we resisted 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 purpose in each language. Then, lemmatized

Table 2 Percentage of the 1,467 lexicon entries found inside the translation memory using several types of query.

Query types	bilingual	English	French
verbatim queries	3 %	9 %	17 %
EN: *I have no axe to grind*			
FR: *Je ne prêche pas pour ma paroisse*			
+ manual removal of extra words	6 %	21 %	28 %
EN: *I have .. axe to grind*			
FR: *Je .. prêche .. pour ma paroisse*			
+ removal of extra pronouns	8 %	30 %	35 %
EN: *have .. axe to grind*			
FR: *prêche .. pour ma paroisse*			
+ verb lemmatization	14 %	43 %	44 %
EN: *have+ .. axe to grind*			
FR: *prêcher+ .. pour ma paroisse*			
+ pronoun and determiner lemmatization	16 %	48 %	48 %
EN: *have+ .. axe to grind*			
FR: *prêcher+ .. pour sa+ paroisse*			

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 thus cannot be replaced by another personal pronoun. As a result, we kept this pronoun verbatim in the queries made. For the same reason, we did not automatically remove negation since it may belong to the idiomatic expression. The idiom "*I did not sleep a wink*" becomes for example incorrect if "*not*" is removed. We are also aware that all verbs or nouns cannot be lemmatized for all idioms, like the verb "*to be*" in "*Enough is enough*". We count on the fact that the incorrect inflection forms of a given expression usually do not occur in the TM.

4.3.1 Observations

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. We know, however, that most queries made by real users of the application do not use those operators. This 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 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 good at identifying the idiomatic expressions we considered. A previous study with this corpus in the medical domain has already shown that the Hansards are a valuable source of information for specialized domains [16]. In this work, we analyzed the responses of TRANSSEARCH with respect to the 20 categories used for labeling the different idiomatic expressions. The main outcomes of this analysis are reported in Table 3. We observe a large discrepancy among classes. While nearly 70 % of expressions in the "Behaving" class were retrieved from the Hansards (e.g. "*go out on a limb for someone*", "*to jump on the bandwagon*"), only 10 % of those belonging to the "Weather" class were found (e.g. "*It's biting cold*", "*The sun is beating down*"). This strengthens the interest of including more bilingual resources inside the TM for better coverage of topics.

Table 3 Coverage of the lexicon entries by the TM of TRANSSEARCH for various topics.

Ranks	English queries		French queries	
1	Behaving	69 %	Behaving	68 %
2	Discussion	68 %	Feelings and emotions	62 %
3	Time, age and experience	65 %	Discussion	61 %
	...		...	
18	Love, sex and seduction	14 %	Human body and physical activity	22 %
19	Weather	14 %	Clothing and fashion	13 %
20	Drinking and eating	8 %	Weather	10 %

5 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. This evaluation encompasses three related experiments: 1) The recall of the trans-

Table 4 Recall (%) measured using the lexicon sanctioned by the translation memory as a reference.

k	1	2	3	5	10	∞
English queries	41.6	56.3	59.2	65.1	69.3	74.8
French queries	41.8	49.8	54.9	62.9	69.6	76.8

lations identified by TRANSSEARCH among the entries of the reference lexicon is first evaluated. 2) These results are then compared with the ones obtained using the widely used and effective MT engine known as GOOGLE TRANSLATE. 3) We finally provide a manual evaluation of the precision of our system.

5.1 Objective Evaluation of the Recall Capabilities of TRANSSEARCH

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 4. 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 (16 %). Therefore we restrained our objective evaluation to those 238 queries. Table 4 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.[9]

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[10] of the query "*be+ around the corner*", most of which being perfectly legitimate translations, while our reference contains only one.

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

[10] 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*.

5.2 Comparison of TRANSSEARCH with GOOGLE TRANSLATE

TRANSSEARCH is able to suggest several translations, provided, however, that enough information is available in its translation memory. We compared its results with those of GOOGLE TRANSLATE online system.[11] As mentioned in the introduction of this article, MT tools do not usually adopt specific strategies for idiomatic expressions. This makes them prone to errors, as was the case in the examples of Table 1. Nevertheless, GOOGLE TRANSLATE has the benefit of relying on bilingual resources that are much broader than those exploited by TRANSSEARCH.

Table 5 shows the recalls measured for both applications. In this case, we took the full lexicon into account. These results were computed for TRANSSEARCH from the queries obtained at the last processing step described in Section 4.3; as far as GOOGLE TRANSLATE is concerned, we took the queries obtained after a manual removal of extra words of the full lexicon entries (step 2 in Table 2) since the MT engine does not have operators equivalent to those of our CAT tool.

The obtained recall values are lower for TRANSSEARCH than for GOOGLE TRANSLATE, particularly at the first rank; this may be explained by the fact that for half of the queries, the concordancer could not find any information in its TM. Surprisingly, the results generated by GOOGLE TRANSLATE are higher than expected. They indicate that a large part of this lexicon is likely to be in the resources used by GOOGLE TRANSLATE and that this system is able to find the corresponding entries inside the translation table.

Table 5 Recall (%) measured taking into account the full lexicon.

	k	1	2	3	5	10
English queries	GOOGLE TRANSLATE	12.3				
	TRANSSEARCH	8.0	10.6	11.3	12.4	13.3
French queries	GOOGLE TRANSLATE	12.6				
	TRANSSEARCH	7.6	9.1	10.3	11.9	13.2

To alleviate the fact that TRANSSEARCH is not able to suggest a translation for all queries in contrast to GOOGLE TRANSLATE, we carried out additional experiments restrained to the queries with at least one result provided by TRANSSEARCH. The findings showed that the recalls (reported in Table 6) are close upon comparing both systems when only the first result displayed by the TM-based system is considered. Finally, TRANSSEARCH suggests several translations, which increases recall from 16 % to 28 %. This is important because it is often the case that a typical user of the concordancer wants to collect different translations of a given expression, something that GOOGLE TRANSLATE does not facilitate.

[11] http://translate.google.com.

Table 6 Recall (%) measured on the 700 English queries and the 705 French queries found respectively in the translation memory of TRANSSEARCH.

	k	1	2	3	5	10
English queries	GOOGLE TRANSLATE	15.7				
	TRANSSEARCH	16.7	22.3	23.6	26.0	27.9
French queries	GOOGLE TRANSLATE	16.7				
	TRANSSEARCH	15.9	18.9	21.4	24.7	27.5

5.3 Manual Evaluation of TRANSSEARCH

While the objective evaluation of the recall capabilities of TRANSSEARCH, presented in Section 5.1, above, revealed the great potential of TRANSSEARCH for translating of idiomatic expressions, it also showed that a manual evaluation of the system was required in order to account for the sparseness of our bilingual lexicon. As a result, we conducted a manual evaluation involving five bilingual annotators who were presented with lists of identified translations for 100 randomly chosen French queries and 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. The annotators were broken up in two groups. The first group consisted of three annotators who judged the first fifty French queries; the second group, consisting of remaning two annotators, judged the next fifty queries.

Across the board the quality appreciated by the annotators turned out to be highly variable, some annotators tending to classify more easily translations as correct. This variability in translation accuracy equated with a low Fleiss inter-annotator agreement value [7] value of 0.25. Figure 4 illustrates some cases of divergence.

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

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

The results of this evaluation are reported in Table 7. To control for the occurrence of inter-rater variability in which a given query can be rated differently by several judges, we decided to credit divergent annotations equally. 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 respectively.

For all but 7 out of 100 queries, TRANSSEARCH was able to identify a translation classified as correct by at least one annotator. For these queries, the average rank of the first correct translation was 1.4. This indicates that relevant translations can usually be found among the first two candidate translations that are displayed by TRANSSEARCH. In addition, on average, we observe that only 36% of the translations proposed to the user are labeled as wrong.

Table 7 Average percentage of translations judged correct, partially correct or wrong per query on a sample of 100 French 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

6 Survey of CAT tools comparable to TRANSSEARCH

This section reviews three recent CAT tools that are able to automatically identify translations from their TM. Similar to TRANSSEARCH, these CAT tools usually resort to statistical word-level alignment methods, bringing them much closer to the capacities of MT engines than classical CAT tools; those last systems being mainly concerned with recycling parallel sentences as a whole.

The CAT systems presented in this section do not use the same techniques to align words found in a given sentence pair, and have different user interfaces as well. Unfortunately, the transpotting methods that are used are seldom described in detail and have not been evaluated with the same kind of rigor (as we showed in our evaluation of TRANSSEARCH) on a significant amount of queries, either with idiomatic expressions such as in this paper or with other expressions considered in a previous study [2]. Experiments should be therefore carried out with the CAT systems to compare them in terms of recall and precision of the identified translations. Since we did not have access to the search engines of these systems, realizing that the lack thereof undoubtly complicates the automatic processing of the results, we decided to focus in this section on the comparison of their functionalities.

6.1 LINEAR B

The CAT tool LINEAR B is available on the Web to translate expressions between English and eight other languages: Arabic, Chinese, Dutch, French, German, Italian, Spanish and Swedish. Figure 5 displays the interface of the system when the query "*is still in its infancy*" was posed to the system. It is reported at the top of the screen

that 18 translations are found. Translations of the query are suggested below, with examples taken from the TM. For long sentences, only a part is displayed, with missing words substituted by '...'. Let us note that at most three occurrences are shown by default for each suggested translation while a hyperlink allows the system access to more examples if they are available inside the TM. For each sentence pair, both the query and its corresponding translation are highlighted in bold.

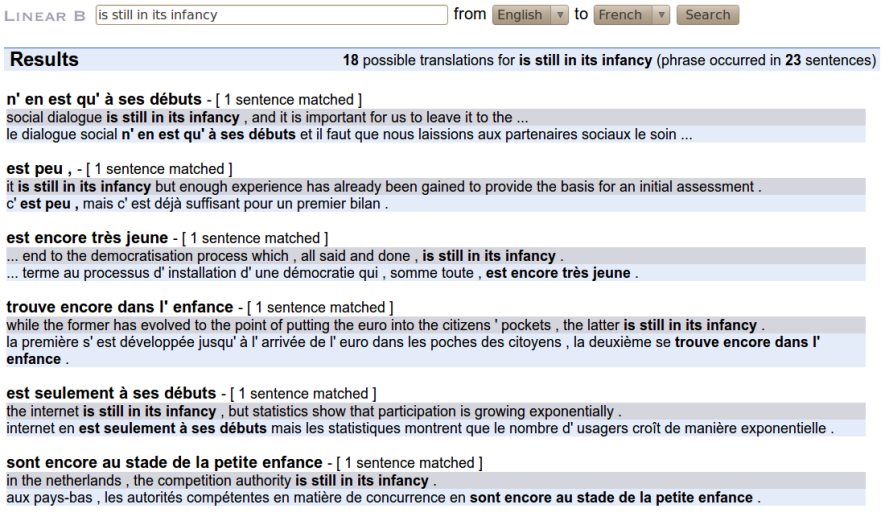

Fig. 5 Result returned by LINEARB to the query "*is still in its infancy*". Only the first six suggested translations are displayed here but the next results can be accessed via a scroll bar.

To the best of our knowledge, it was the first online CAT system which had the capacity to automatically identify translations of expressions using recent SMT methods. The process that builds the TM and then efficiently searches translations is described in [4]. In short, a phrase table is automatically built from parallel corpora aligned at the sentence level drawing on methods that are usually employed to train SMT models [8]. This phrase table is then stored in a suffix array data structure in order to efficiently look up the possible translations of a phrase.

For expressions that are not found verbatim in the TM, the system provides the list of the subsequences (sequences that are derived from another sequence) that were found inside the indexed phrase table. For example, the interface reports that a partial match was found for the query "*proud as a peacock*" when translating to French and suggests three subsets of queries (Fig. 6). It is important to note that no subsequence with the term "*peacock*" was found inside the TM.

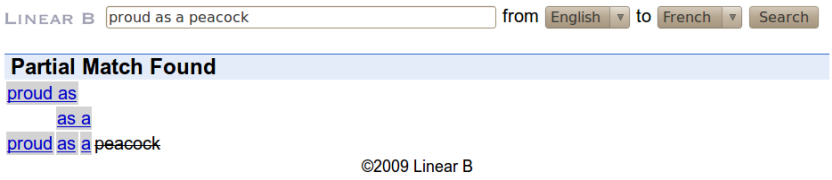

Fig. 6 Result returned by LINEARB to the query "*proud as a peacock*".

6.2 LINGUEE

LINGUEE is a CAT tool combining a dictionary and a search engine with which users can search through a TM for words and expressions. Developed by Linguee GmbH, the system was officially launched online in 2010[12] to translate between English and 4 other languages: French, German, Portuguese and Spanish. According to a press article released in 2011 by the company, LINGUEE answers 1.5 million search requests every day.

Figure 7 shows the default view for the query "*is still in its infancy*". Vocabulary entries from the dictionary is displayed on the left. If the queried expression had been found in the dictionary, the translation would have been shown on the top of the screen. The dictionary entries that match partially the expression are displayed below. On the bottom — not displayed in Figure 7, a button "Suggest translation" allows users to type a translation for the query; this translation will be added to the dictionary after being checked by an editor. The right hand side displays example sentences from the TM. For each sentence pair, its origin is shown (the Europarl corpus for the first sentence of our example), while the query and its hypothesized translation are highlighted. The full text that contains the sentence pair can be downloaded by clicking on the hyperlink displaying the origin. Various color intensities show the confidence in the alignment of a word of the hypothesized translation with the query. Like LINEAR B, long sentences are cut, with missing parts replaced by '[...]'. The sentence pairs are not organized according to the identified translations. The user can rate a translation by clicking on the thumbs symbol which appears on the right side of each sentence pair. By rating translations, the user can influence the order in which the search results are displayed following future search requests.

According to the LINGUEE website, the TM is made of hundreds of millions of bilingual texts. The majority of the indexed sentence pairs are from professionally translated websites of organizations, companies and universities. Other sources include European Union documents and patent specifications, as well as various Web pages crawled and filtered with a machine-learning model. This model is continuously trained using the user feedback on the translation quality. Since some sources are less trustworthy, a warning sign is added before some displayed pair of sentences in order to indicate a possibly wrong translation. The highlighting of the identified

[12] http://www.linguee.com.

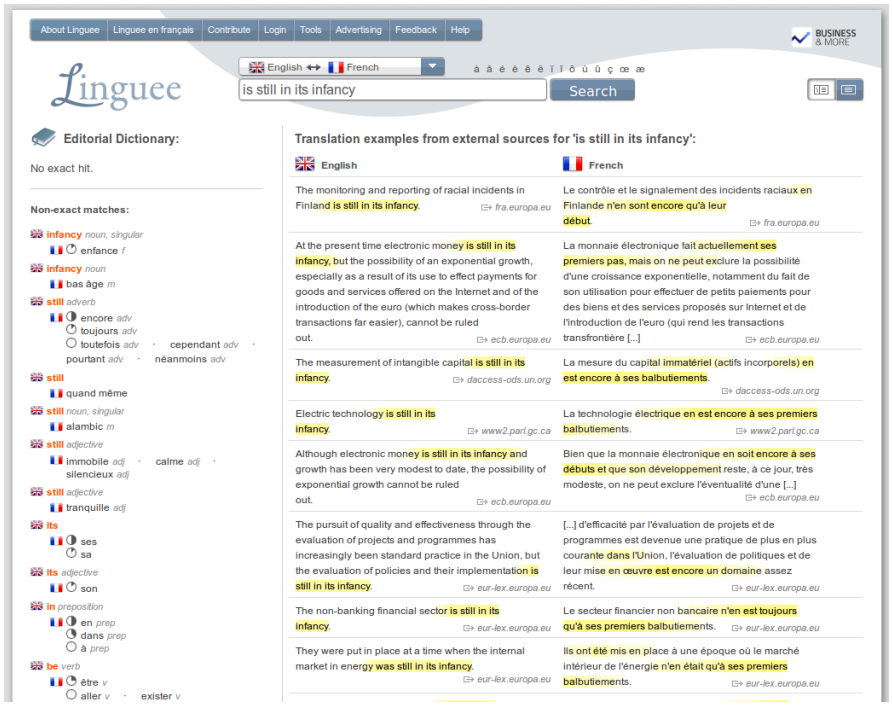

Fig. 7 Result returned by LINGUEE to the query "*is still in its infancy*".

translation in the example sentences is generated automatically though we were not able to find information on this process; word-based algorithms [3] used in MT are likely to be used for this purpose.

The query submission system integrates different functionalities. A "did you mean" feature suggests an expression close to the submitted query when the number of hits in the TM is too small. Possible completions of queries are also displayed when typing. Quotes can be used to search for exact phrases in order to find only sentences in which the query words occur in the exact form and order. Two operators are introduced: a '-' sign is helpful to exclude words; a '+' sign forces the inclusion of words in a certain form, which is helpful with pronouns which would otherwise be ignored by LINGUEE. Finally a morphological analysis enables the system to search some inflected forms of a given query. For example, sentence pairs containing "*was still in its infancy*" or "*is still in its infancy*" are retrieved for the query "*being still in its infancy*".

6.3 TRADOOIT

TRADOOIT is a CAT developed by Okidoo Inc. It includes a TM, a term bank and a bilingual concordancer. This product has been marketed since 2008 and a free version is available online since November 2011 for the English-French pair.[13] Only this free version is reviewed here.

The system uses two kinds of resources. The first one is a TM made of various parallel corpora, delivered mainly by the Canadian government (including the Hansards), the Ontario government, the European Parliament, the World Trade Organization, UNESCO and opensubtitles[14] which provides movie subtitles. The whole TM contains more than 200 M words per language. The second resource type includes various term banks, including Wikipedia, and TERMIUM PLUS[15] which is the Government of Canada's terminology and linguistic data bank. These two kinds of resources are searched at the same time when a query is submitted. In the displayed example of Figure 8, information was only found in the parallel corpora but the opposite situation may happen. For example, the query "*is hard of hearings*" does not have a single hit in the TM but translations of the related expressions "*person who is hard of hearing*" and "*employee who is hard of hearing*" are shown from the TERMIUM PLUS bank.

The system interface displays on the left hand side various information (Fig. 8). The "Grouped Translations" section displays statistics on the different translations identified in the TM with their frequency. The "1001 Forms" section lists the various forms found for the searched expression, i.e. forms that differ on capitalization or inflection (conjugation or plural when the canonical form of verbs and nouns is searched). The "Sources" section allows users to refine their search by filtering results based on the corpus origin. The right hand side displays TM sentence pairs in table format and highlights source language hits and their target language equivalent. The translation of the query is automatically identified inside the sentence pairs using a method that is not described. For each use example, the source is specified and the hyperlink "See bitext" allows users to access the sentence in the context of the source document. Two thumb pictures are also displayed for each sentence pair. This enables users to provide a feedback on the quality of the identified translation.

The concordancer provides additional features for query processing. For example, the system is able to suggest searches that may be more successful in case the user gets too few results. This process of suggesting alternate searches also includes the use of several operators: 1) a '+' sign added at the end of a word allows search on the various inflected forms of this word; 2) a '?' sign indicates that a word is missing in the typed expression at a given position; and 3) a '*' sign means that zero or one word can occur at a given position of the expression. Since these various signs can be combined, this allows the user, for example, to search together "*write a letter*", "*writing letters*" or "*write detailed letters*" with the query "*write+ ? letter+*".

[13] http://www.tradooit.com.

[14] http://www.opensubtitles.org.

[15] http://www.termiumplus.gc.ca.

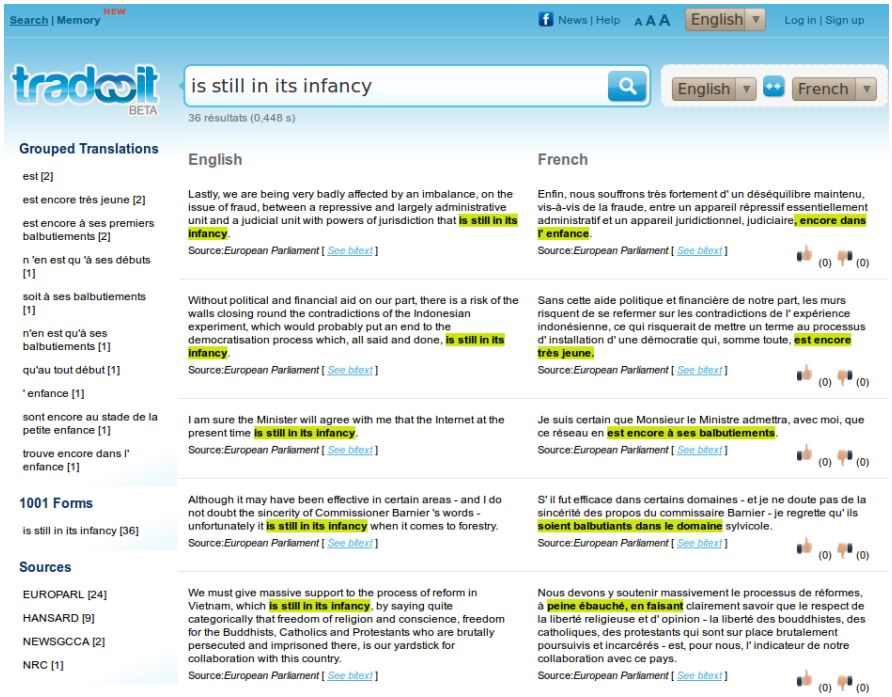

Fig. 8 Result returned by TRADOOIT to the query "*is still in its infancy*".

7 Conclusion

In this work, we have studied the problem of identifying translations of idiomatic expressions in both English and French, with a brand new version of the bilingual concordancer TRANSSEARCH. We showed in our experiments that a user who would query the system verbatim would often fail to find a match in the TM. As a result, some innovation is required in order to get good use of the system, such as utilizing the morphological ('+') and the proximity ('..') operators available in the query language recognized by the system. We automatized the querying process and conducted experiments that search entries of a phrase book inside a TM collected from the Canadian Hansards. These experiments showed that almost half of the 1.5 thousand idiomatic expressions queried to the system finally got a match in the TM, while a high proportion of the translations returned by the automated system were correct.

A comparison of the output generated by TRANSSEARCH with the GOOGLE TRANSLATE MT system showed that in spite of a relatively small size of its TM, our concordancer has a decent recall. And in fact, it even obtained higher recall values than GOOGLE TRANSLATE if several translations are considered, which is the typical modus operandi of our application.

Finally, we also discussed the functionalities of recent substantial bilingual concordancers. While an evaluation of those systems for identifying idiomatic expressions was out of the scope of the present study, the existence of several such applications shows the increasing popularity of advanced TM systems in the sphere of professional translators. In fact, the TRANSSEARCH application we used in this research is now released by Terminotix.[16]

Acknowledgments

This work was funded by an NSERC grant in collaboration with Terminotix.[17] We are indebted to Sandy Dincky, Fabienne Venant and Neil Stewart who kindly participated to the annotation task.

References

1. Dimitra Anastasiou. Identification of idioms by machine translation: a hybrid research system vs. three commercial systems. In *Proceedings of EAMT*, pages 12–20, Hamburg, Germany, 2008.
2. Julien Bourdaillet, Stéphane Huet, Philippe Langlais, and Guy Lapalme. TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation*, 24(3–4):241–271, 2010.
3. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):2, 1993.
4. Chris Callison-Burch, Colin Bannard, and Josh Shroeder. A compact data structure for searchable translation memories. In *Proceedings of EAMT*, pages 59–65, Budapest, Hungary, 2005.
5. 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.
6. Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103, 2009.
7. 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.
8. Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, volume 1, pages 48–54, Edmonton, Canada, 2003.
9. 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.
10. Philippe Langlais. A system to align complex bilingual corpora. Technical report, CTT, KTH, Stockholm, Sweden, 1997.
11. Elliott Macklovitch, Guy Lapalme, and Fabrizio Gotti. TransSearch: What are translators looking for? In *Proceedings of AMTA*, pages 412–419, Waikiki, Hawaii, USA, 2008.
12. 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.

[16] http://www.tsrali3.com.

[17] http:/www.terminotix.com.

13. Tom McArthur, editor. *The Oxford Companion to the English Language*. Oxford University Press, 1992.
14. Igor Mel'čuk. *Idioms: Structural and Psychological Perspectives*, chapter Phrasemes in Language and Phraseology in Linguistics, pages 167–232. Hillsdale, NJ: Lawrence Erlbaum, 1995.
15. Igor Mel'čuk. La phraséologie en langue, en dictionnaire et en TALN. In *Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Montreal, Canada, 2010.
16. Aurélie Névéol and Sylwia Ozdowska. Terminologie médicale bilingue anglais/français : usages clinique et législatif. *Glottopol*, 8:5–21, 2006.
17. Jean-Bernard Piat. *It's raining cats and dogs et autres expressions idiomatiques anglaises*. Librio. J'ai lu, 2008.
18. Alain Polguère. *Lexicologie et sémantique lexicale : notions fondamentales*. Les Presses de l'Université de Montréal, 2nd edition, 2008.
19. 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.
20. 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.
21. 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.
22. Koichi Takeuchi, Takashi Kanehila, Kazuki Hilao, Takeshi Abekawa, and Kyo Kageura. Flexible automatic look-up of English idiom entries in dictionaries. In *Proceedings of MT Summit*, pages 451–458, Copenhagen, Denmark, 2007.
23. Jean Véronis and Philippe Langlais. *Evaluation of Parallel Text Alignment Systems — The Arcade Project.*, chapter 19, pages 369–388. Kluwer Academic Publisher, 2000.
24. Martin Volk. *Machine Translation: Theory, Applications, and Evaluation. An Assessment of the State-of-the-Art*, chapter The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems, pages 167–192. Gardez! Verlag, St. Augustin, Germany, 1998.