Identifying idiomatic expressions using automatic word-alignment

Begoña Villada Moirón and Jörg Tiedemann
Alfa Informatica, University of Groningen
Oude Kijk in 't Jatstraat 26
9712 EK Groningen, The Netherlands
{M.B.Villada.Moiron,J.Tiedemann}@rug.nl

Abstract

For NLP applications that require some sort of semantic interpretation it would be helpful to know what expressions exhibit an idiomatic meaning and what expressions exhibit a literal meaning. We investigate whether automatic word-alignment in existing parallel corpora facilitates the classification of candidate expressions along a continuum ranging from literal and transparent expressions to idiomatic and opaque expressions. Our method relies on two criteria: (i) meaning predictability that is measured as semantic entropy and (ii), the overlap between the meaning of an expression and the meaning of its component words. We approximate the mentioned overlap as the proportion of default alignments. We obtain a significant improvement over the baseline with both measures.

1 Introduction

Knowing whether an expression receives a literal meaning or an idiomatic meaning is important for natural language processing applications that require some sort of semantic interpretation. Some applications that would benefit from knowing this distinction are machine translation (Imamura et al., 2003), finding paraphrases (Bannard and Callison-Burch, 2005), (multilingual) information retrieval (Melamed, 1997a), etc.

The purpose of this paper is to explore to what extent word-alignment in parallel corpora can be used to distinguish idiomatic multiword expressions from more transparent multiword expressions and fully productive expressions.

In the remainder of this section, we present our characterization of idiomatic expressions, the motivation to use parallel corpora and related work. Section 2 describes the materials required to apply our method. Section 3 portraits the routine to extract a list of candidate expressions from automatically annotated data. Experiments with different word alignment types and metrics are shown in section 4. Our results are discussed in section 5. Finally, we draw some conclusions in section 6.

1.1 What are idiomatic expressions?

Idiomatic expressions constitute a subset of multiword expressions (Sag et al., 2001). We assume that literal expressions can be distinguished from idiomatic expressions provided we know how their meaning is derived.[1] The meaning of linguistic expressions can be described within a scale that ranges from fully transparent to opaque (in figurative expressions).

(1) Wat moeten lidstaten ondernemen om
what must member states do to
aan haar eisen te voldoen?
at her demands to meet?
'What must EU member states do to meet her demands?'

(2) Deze situatie **brengt** de bestaande politieke
this situation brings the existing political
barrières zeer duidelijk **aan het licht**.
barriers very clearly in the light
'This situation brings the existing political limitations to light very clearly.'

[1] Here, we ignore morpho-syntactic and pragmatic factors that could help model the distinction.

(3) Wij mogen ons hier niet bij neerleggen,
we may us here not by agree,
maar moeten de situatie publiekelijk **aan**
but must the situation publicly op
de kaak stellen.
the cheek state
'We cannot agree but must denounce the situation openly.'

Literal and transparent meaning is associated with high meaning predictability. The meaning of an expression is fully predictable if it results from combining the meaning of its individual words when they occur in isolation (see (1)). When the expression undergoes a process of metaphorical interpretation its meaning is less predictable. Moon (1998) considers a continuum of transparent, semi-transparent and opaque metaphors. The more transparent metaphors have a rather predictable meaning (2); the more opaque have an unpredictable meaning (3). In general, an unpredictable meaning results from the fact that the meaning of the expression has been fossilized and conventionalized. In an uninformative context, idiomatic expressions have an unpredictable meaning (3). Put differently, the meaning of an idiomatic expression cannot be derived from the cumulative meaning of its constituent parts when they appear in isolation.

1.2 Why checking translations?

This paper addresses the task of distinguishing literal (transparent) expressions from idiomatic expressions. Deciding what sort of meaning an expression shows can be done in two ways:

- measuring how predictable the meaning of the expression is and
- assessing the link between (a) the meaning of the expression as a whole and (b) the cumulative literal meanings of the components.

Fernando and Flavell (1981) observe that no connection between (a) and (b) suggests the existence of opaque idioms and, a clear link between (a) and (b) is observed in clearly perceived metaphors and literal expressions.

We believe we can approximate the meaning of an expression by looking up the expressions' translation in a foreign language. Thus, we are interested in exploring to what extent parallel corpora can help us to find out the type of meaning an expression has.

For our approach we make the following assumptions:

- regular words are translated (more or less) consistently, i.e. there will be one or only a few highly frequent translations whereas translation alternatives will be infrequent;
- an expression has a (almost) literal meaning if its translation(s) into a foreign language is the result of combining each word's translation(s) when they occur in isolation into a foreign language;
- an expression has a non-compositional meaning if its translation(s) into a foreign language does not result from a combination of the regular translations of its component words.

We also assume that an automatic word aligner will get into trouble when trying to align non-decomposable idiomatic expressions word by word. We expect the aligner to produce a large variety of links for each component word in such expressions and that these links are different from the default alignments found in the corpus otherwise.

Bearing these assumptions in mind, our approach attempts to locate the translation of a MWE in a target language. On the basis of all reconstructed translations of a (potential) MWE, it is decided whether the original expression (in source language) is idiomatic or a more transparent one.

1.3 Related work

Melamed (1997b) measures the semantic entropy of words using bitexts. Melamed computes the translational distribution T of a word s in a source language and uses it to measure the translational entropy of the word $H(T|s)$; this entropy approximates the semantic entropy of the word that can be interpreted either as (a) the semantic ambiguity or (b) the inverse of reliability. Thus, a word with high semantic entropy is potentially very ambiguous and therefore, its translations are less reliable (or highly context-dependent). We also use entropy to approximate meaning predictability. Melamed (1997a) investigates various techniques to identify non-compositional compounds in parallel data. Non-compositional compounds

are those sequences of 2 or more words (adjacent or separate) that show a conventionalized meaning. From English-French parallel corpora, Melamed's method induces and compares pairs of translation models. Models that take into account non-compositional compounds are highly accurate in the identification task.

2 Data and resources

We base our investigations on the Europarl corpus consisting of several years of proceedings from the European Parliament (Koehn, 2003). We focus on Dutch expressions and their translations into English, Spanish and German.[2] Thus, we used the entire sections of Europarl in these three languages. The corpus has been tokenized and aligned at the sentence level (Tiedemann and Nygaard, 2004). The Dutch part contains about 29 million tokens in about 1.2 million sentences. The English, Spanish and German counterparts are of similar size between 28 and 30 million words in roughly the same number of sentences.

Automatic word alignment has been done using GIZA++ (Och, 2003). We used standard settings of the system to produce Viterbi alignments of IBM model 4. Alignments have been produced for both translation directions (source to target and target to source) on tokenized plain text.[3] We also used a well-known heuristics for combining the two directional alignments, the so-called refined alignment (Och et al., 1999). Word-to-word alignments have been merged such that words are connected with each other if they are linked to the same target. In this way we obtained three different word alignment files: source to target (*src2trg*) with possible multi-word units in the source language, target to source (*trg2src*) with possible multi-word units in the target language, and *refined* with possible multi-word units in both languages. We also created bilingual word type links from the different word-aligned corpora. These lists include alignment frequencies that we will use later on for extracting default alignments for individual words. Henceforth, we will call them *link lexica*.

[2]This is only a restriction for our investigation but not for the approach itself.

[3]Manual corrections and evaluations of the tokenization, sentence and word alignment have not been done. We rely entirely on the results of automatic processes.

3 Extracting candidates from corpora

The Dutch section from the Europarl corpus was automatically parsed with Alpino, a Dutch wide-coverage parser.[4] 1.25% of the sentences could not be parsed by Alpino, given the fact that many sentences are rather lengthy. We selected those sentences in the Dutch Europarl section that contain at least one of a group of verbs that can function as main or support verbs. Support verbs are prone to lexicalization or idiomatization along with their complementation (Butt, 2003). The selected verbs are: *doen*, *gaan*, *geven*, *hebben*, *komen*, *maken*, *nemen*, *brengen*, *houden*, *krijgen*, *stellen* and *zitten*.[5]

A fully parsed sentence is represented by the list of its dependency triples. From the dependency triples, each main verb is tallied with every dependent prepositional phrase (PP). In this way, we collected all the VERB PP tuples found in the selected documents. To avoid data sparseness, the NP inside the PP is reduced to the head noun's lemma and verbs are lemmatized, too. Other potential arguments under a verb phrase node are ignored. A sample of more than 191,000 candidates types (413,000 tokens) was collected. To ensure statistical significance, the types that occur less than 50 times were ignored.

For each candidate triple, the log-likelihood (Dunning, 1993) and salience (Kilgarriff and Tugwell, 2001) scores were calculated. These scores have been shown to perform reasonably well in identifying collocations and other lexicalized expressions (Villada Moirón, 2005). In addition, the head dependence between each PP in the candidates dataset and its selecting verbs was measured. Merlo and Leybold (2001) used the head dependence as a diagnostic to determine the argument (or adjunct) status of a PP. The head dependence is measured as the amount of entropy observed among the co-occurring verbs for a given PP as suggested in (Merlo and Leybold, 2001; Baldwin, 2005). Using the two association measures and the head dependence heuristic, three different rankings of the candidate triples were produced. The three different ranks assigned to each triple were uniformly combined to form the final ranking. From this list, we selected the top 200 triples

[4]Available at `http://www.let.rug.nl/~vannoord/alp/Alpino`.

[5]Butt (2003) maintains that the first 7 verbs are examples of support verbs crosslinguistically. The other 5 have been suggested for Dutch by (Hollebrandse, 1993).

which we considered a manageable size to test our method.

4 Methodology

We examine how expressions in the source language (Dutch) are conceptualized in a target language. The translations in the target language encode the meaning of the expression in the source language. Using the translation links in parallel corpora, we attempt to establish what type of meaning the expression in the source language has. To accomplish this we make use of the three word-aligned parallel corpora from Europarl as described in section 2.

Once the translation links of each expression in the source language have been collected, the entropy observed among the translation links is computed per expression. We also take into account how often the translation of an expression is made out of the default alignment for each triple component. The default 'translation' is extracted from the corresponding bilingual link lexicon.

4.1 Collecting alignments

For each triple in the source language (Dutch) we collect its corresponding (hypothetical) translations in a target language. Thus, we have a list of 200 VERB PP triples representing 200 potential MWEs in Dutch. We selected all occurrences of each triple in the source language and all aligned sentences containing their corresponding translations into English, German and Spanish. We restricted ourselves to instances found in 1:1 sentence alignments. Other units contain many errors in word and sentence alignment and, therefore, we discarded them. Relying on automated word-alignment, we collect all translation links for each verb, preposition and noun occurrence within the triple context in the three target languages.

To capture the meaning of a source expression (triple) S, we collect all the translation links of its component words s in each target language. Thus, for each triple, we gather three lists of translation links T_s. Let us see the example AAN LICHT BRENG representing the MWE *iets aan het licht brengen* 'reveal'. Table 1 shows some of the links found for the triple AAN LICHT BRENG. If a word in the source language has no link in the target language (which is usually due to alignments to the empty word), NO_LINK is assigned.

Note that Dutch word order is more flexible than English word order and that, the PP argument in a candidate expression may be separate from its selecting verb by any number of constituents. This introduces much noise during retrieving translation links. In addition, it is known that concepts may be lexicalized very differently in different languages. Because of this, words in the source language may translate to nothing in a target language. This introduces many mappings of a word to NO_LINK.

Triple	Links in English
aan	NO_LINK, to, of, in, for, from, on, into, at
licht	NO_LINK, light, revealed, exposed, highlight, shown, shed light, clarify
breng	NO_LINK, brought, bring, highlighted, has, is, makes

Table 1: Excerpt of the English links found for the triple AAN LICHT BRENG 'bring to light'.

4.2 Measuring translational entropy

According to our intuition it is harder to align words in idiomatic expressions than other words. Thus, we expect a larger variety of links (including erroneous alignments) for words in such expressions than for words taken from expressions with a more literal meaning. For the latter, we expect fewer alignment candidates, possibly with only one dominant default translation. Entropy is a good measure for the unpredictability of an event. We like to use this measure for comparing the alignment of our candidates and expect a high average entropy for idiomatic expressions. In this way we approximate a measure for meaning predictability.

For each word in a triple, we compute the entropy of the aligned target words as shown in equation (1).

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) log P(t|s) \qquad (1)$$

This measure is equivalent to translational entropy (Melamed, 1997b). $P(t|s)$ is estimated as the proportion of alignment t among all alignments of word s found in the corpus in the context of the given triple.[6] Finally, the translational entropy of a triple is the average translational entropy of its components. It is unclear how to

[6]Note that we also consider cases where s is part of an aligned multi-word unit.

treat NO_LINKS. Thus, we experiment with three variants of entropy: (1) leaving out NO_LINKS, (2) counting NO_LINKS as multiple types and (3) counting all NO_LINKS as one unique type.

4.3 Proportion of default alignments (pda)

If an expression has a literal meaning, we expect the default alignments to be accurate literal translations. If an expression has idiomatic meaning, the default alignments will be very different from the links observed in the translations.

For each triple S, we count how often each of its components s is linked to one of the default alignments D_s. For the latter, we used the four most frequent alignment types extracted from the corresponding link lexicon as described in section 2. A large proportion of default alignments[7] suggests that the expression is very likely to have literal meaning; a low percentage is suggestive of non-transparent meaning. Formally, pda is calculated in the following way:

$$pda(S) = \frac{\sum_{s \in S} \sum_{d \in D_s} align\_freq(s,d)}{\sum_{s \in S} \sum_{t \in T_s} align\_freq(s,t)} \quad (2)$$

where $align\_freq(s,t)$ is the alignment frequency of word s to word t in the context of the triple S.

5 Discussion of experiments and results

We experimented with the three word-alignment types (src2trg, trg2src and refined) and the two scoring methods (entropy and pda). The 200 candidate MWEs have been assessed and classified into idiomatic or literal expressions by a human expert. For assessing performance, standard precision and recall are not applicable in our case because we do not want to define an artificial cut-off for our ranked list but evaluate the ranking itself. Instead, we measured the performance of each alignment type and scoring method by obtaining another evaluation metric employed in information retrieval, *uninterpolated average precision* (uap), that aggregates precision points into one evaluation figure. At each point c where a true positive S_c in the retrieved list is found, the precision $P(S_1..S_c)$ is computed and, all precision points are then averaged (Manning and Schütze, 1999).

[7] Note that we take NO_LINKS into account when computing the proportions.

$$uap = \frac{\sum_{S_c} P(S_1..S_c)}{|S_c|} \quad (3)$$

We used the initial ranking of our candidates as baseline. Our list of potential MWEs shows an overall precision of 0.64 and an uap of 0.755.

5.1 Comparing word alignment types

Table 2 summarizes the results of using the entropy measure (leaving out NO_LINKS) with the three alignment types for the NL-EN language pair.[8]

Alignment	uap
src2trg	0.864
trg2src	0.785
refined	0.765
baseline	0.755

Table 2: uap values of various alignments.

Using word alignments improves the ranking of candidates in all three cases. Among them, src2trg shows the best performance. This is surprising because the quality of word-alignment from English-to-Dutch (trg2src) in general is higher due to differences in compounding in the two languages. However, this is mainly an issue for noun phrases which make up only one component in the triples.

We assume that src2trg works better in our case because in this alignment model we explicitly link each word in the source language to exactly one target word (or the empty word) whereas in the trg2src model we often get multiple words (in the target language) aligned to individual words in the triple. Many errors are introduced in such alignment units. Table 3 illustrates this with an example with links for the Dutch triple *op prijs stel* corresponding to the expression *iets op prijs stellen* 'to appreciate sth.'

src2trg		trg2src	
source	target	target	source
gesteld	appreciate	NO_LINK	stellen
prijs	appreciate	much appreciate indeed	prijs
op	appreciate	NO_LINK	op
gesteld	be	keenly appreciate	stellen
prijs	delighted	fact	prijs
op	NO_LINK	NO_LINK	op

Table 3: Example src2trg and trg2src alignments for the triple OP PRIJS STEL.

[8] The performance of the three alignment types remains uniform across all chosen language pairs.

src2trg alignment proposes *appreciate* as a link to all three triple components. This type of alignment is not possible in trg2src. Instead, trg2src includes two NO_LINKS in the first example in table 3. Furthermore, we get several multiword-units in the target language linked to the triple components also because of alignment errors. This way, we end up with many NO_LINKS and many alignment alternatives in trg2src that influence our entropy scores. This can be observed for idiomatic expressions as well as for literal expressions which makes translational entropy less reliable in trg2src alignments for contrasting these two types of expressions.

The *refined* alignment model starts with the intersection of the two directional models and adds iteratively links if they meet some adjacency constraints. This results in many NO_LINKS and also alignments with multiple words on both sides. This seems to have the same negative effect as in the trg2src model.

5.2 Comparing scoring metrics

Table 4 offers a comparison of applying translational entropy and the *pda* across the three language pairs. To produce these results, `src2trg` alignment was used given that it reaches the best performance (refer to Table 2).

Score	NL-EN	NL-ES	NL-DE
entropy			
- without NO_LINKS	0.864	0.892	0.907
- NO_LINKS=many	0.858	0.890	0.883
- NO_LINKS=one	0.859	0.890	0.911
pda	0.891	0.894	0.894
baseline	0.755	0.755	0.755

Table 4: Translational entropy and the pda across three language pairs. Alignment is `src2trg`.

All scores produce better rankings than the baseline. In general, pda achieves a slightly better accuracy than entropy except for the NL-DE language pair. Nevertheless, the difference between the metrics is hardly significant.

5.3 Further improvements

One problem in our data is that we deal with wordform alignments and not with lemmatized versions. For Dutch, we know the lemma of each word instance from our candidate set. However, for the target languages, we only have access to surface forms from the corpus. Naturally, inflectional variations influence entropy scores (because of the larger variety of alignment types) and also the pda scores (where the exact wordforms have to be matched with the default alignments instead of lemmas). In order to test the effect of lemmatization on different language pairs, we used CELEX (Baayen et al., 1993) for English and German to reduce wordforms in the alignments and in the link lexicon to corresponding lemmas. We assigned the most frequent lemma to ambiguous wordforms. Table 5 shows the scores obtained from applying lemmatization for the `src2trg` alignment using entropy (without NO_LINKS) and pda.

Setting	NL-EN	NL-ES	NL-DE
using entropy scores			
with prepositions			
wordforms	0.864	0.892	0.907
lemmas	0.873	–	0.906
without prepositions			
wordforms	0.906	0.923	**0.932**
lemmas	0.910	–	0.931
using pda scores			
with prepositions			
wordforms	0.891	0.894	0.894
lemmas	0.888	–	0.903
without prepositions			
wordforms	0.897	0.917	0.905
lemmas	0.900	–	0.910
baseline	0.755	0.755	0.755

Table 5: Translational entropy and pda from `src2trg` alignments across languages pairs with different settings.

Surprisingly, lemmatization adds little or even decreases the accuracy of the pda and entropy scores. It is also surprising that lemmatization does not affect the scores for morphologically richer languages such as German (compared to English). One possible reason for this is that lemmatization discards morphological information that is crucial to identify idiomatic expressions. In fact, nouns in idiomatic expressions are more fixed than nouns in literal expressions. By contrast, verbs in idiomatic expressions often allow tense inflection. By clustering wordforms into lemmas we lose this information. In future work, we might lemmatize only the verb.

Another issue is the reliability of the word alignment that we base our investigation upon. We want to make use of the fact that automatic word alignment has problems with the alignment of individual words that belong to larger lexical units. However, we believe that the alignment program in general has problems with highly ambiguous words such as prepositions. Therefore, preposi-

tions might blur the contrast between idiomatic expressions and literal translations when measured on the alignment of individual words. Table 5 includes scores for ranking our candidate expressions with and without prepositions. We observe that there is a large improvement when leaving out the alignments of prepositions. This is consistent for all language pairs and the scores we used for ranking.

rank	pda	entropy	MWE	triple
1	9.80	8.3585	ok	breng tot stand 'create'
2	9.24	8.0923	ok	breng naar voren 'bring up'
3	16.40	7.8741	ok	kom in aanmerking 'qualify'
4	15.33	7.8426	ok	kom tot stand 'come about'
5	8.70	7.4973	ok	stel aan orde 'bring under discussion'
6	5.65	7.4661	ok	ga te werk 'act unfairly'
7	17.46	7.4057	ok	kom aan bod 'get a chance'
8	9.38	7.1762	ok	ga van start 'proceed'
9	14.15	7.1009	ok	stel aan kaak 'expose'
10	18.75	7.0321	ok	breng op gang 'get going'
11	13.00	6.9304	ok	kom ten goede 'benefit'
12	1.78	6.8715	ok	neem voor rekening 'pay costs'
13	20.99	6.7411	ok	kom tot uiting 'manifest'
14	1.41	6.7360	ok	houd in stand 'preserve'
15	0.81	6.6426	ok	breng in kaart 'chart'
16	16.71	6.5194	ok	breng onder aandacht 'bring to attention'
17	10.25	6.4893	ok	neem onder loep 'scrutinize'
18	7.83	6.4666	ok	breng aan licht 'reveal'
19	5.99	6.4049	ok	roep in leven 'set up'
20	15.89	6.3729	ok	neem in aanmerking 'consider'
...				
100	1.72	4.6940	ok	leg aan band 'control'
101	14.91	4.6884	ok	houd voor gek 'pull s.o.'s leg'
102	23.56	4.6865	ok	kom te weten 'find out'
103	15.38	4.6713	ok	neem in ontvangst 'receive'
104	31.57	4.6556	*	ga om waar 'go about where'
105	35.95	4.6380	*	houd met daar 'keep with there'
106	34.86	4.6215	*	ga om zaak 'go about issue'
107	28.33	4.5846	ok	kom tot overeenstemming 'come to terms'
108	6.06	4.5715	ok	breng in handel 'launch'
109	35.62	4.5370	*	ga om bedrag 'go about amount'
110	22.58	4.5089	*	blijk uit feit 'seems from fact'
111	51.12	4.4063	ok	ben van belang 'matter'
112	49.69	4.3921	*	ga om kwestie 'go about issue'
113	23.61	4.3902	*	voorzie in behoefte 'fill gap'
114	16.18	4.3568	ok	geef aan oproep 'make appeal'
115	50.00	4.3254	*	houd met aspect 'keep with aspect'
116	40.91	4.3006	*	houd aan regel 'adhere to rule'
117	20.12	4.3002	*	stel_vast met voldoening 'settle with satisfaction'
118	36.90	4.2931	ok	kom tot akkoord 'reach agreement'
119	36.49	4.2906	ok	breng in stemming 'get in mood'
120	14.06	4.2873	ok	sta op schroeven 'be unsettled'
...				
180	70.53	2.7395	*	voldoe aan criterium 'satisfy criterion'
181	52.33	2.7351	*	beschik over informatie 'decide over information'
182	74.71	2.6896	*	stem voor amendement 'vote for amending'
183	76.56	2.5883	*	neem_deel aan stemming 'participate in voting'
184	30.26	2.4484	ok	kan op aan 'be able to trust'
185	68.89	2.3199	*	zeg tegen heer 'tell a gentleman'
186	45.00	2.1113	*	verwijs_terug naar commissie 'refer to comission'
187	80.39	2.0992	*	stem tegen amendement 'vote againsta amending'
188	78.04	2.0924	*	onthoud van stemming 'withhold one's vote'
189	77.63	1.9997	*	feliciteer met werk 'congratulate with work'
190	82.21	1.9020	*	stem voor verslag 'vote for report'
191	77.78	1.9016	*	schep van werkgelegenheid 'set up of employment'
192	86.36	1.8775	*	stem voor resolutie 'vote for resolution '
193	73.33	1.8687	*	bedank voor feit 'thank for fact'
194	39.13	1.8497	*	was_wit van geld 'wash money'
195	82.20	1.7944	*	stem tegen verslag 'vote against report'
196	80.49	1.6443	*	schep van baan 'set up of job'
197	86.17	1.4260	*	stem tegen resolutie 'vote against resolution'
198	85.56	1.1779	*	dank voor antwoord 'thank for reply'
199	90.55	1.0398	*	ontvang overeenkomstig artikel 'receive similar article'
200	87.88	1.0258	*	recht van vrouw 'right of woman'

Table 6: Rank (using entropy), entropy score, and pda score of 60 candidate MWEs.

Table 6 provides an excerpt from the ranked list of candidate triples. The ranking has been done using src2trg alignments from Dutch to German with the best setting (see table 5). The score assigned by the pda metric is also shown. The column labeled MWE states whether the expression is idiomatic ('ok') or literal ('*'). One issue that emerges is whether we can find a threshold value that splits candidate expressions into idiomatic and transparent ones. One should choose such a threshold empirically however, it will depend on what level of precision is desirable and also on the final application of the list.

6 Conclusion and future work

In this paper we have shown that assessing automatic word alignment can help to identify idiomatic multi-word expressions. We ranked candidates according to their link variability using translational entropy and their link consistency with regards to default alignments. For our experiments we used a set of 200 Dutch MWE candidates and word-aligned parallel corpora from Dutch to English, Spanish and German. The MWE candidates have been extracted using standard association measures and a head dependence heuristic. The word alignment has been done using standard models derived from statistical machine translation. Two measures were tested to re-rank the candidates. Translational entropy measures the predictability of the translation of an expression by looking at the links of its components to a target language. Ranking our 200 MWE candidates using entropy on Dutch to German word alignments improved the baseline of 75.5% to 93.2% uninterpolated average precision (uap). The proportion of default alignments among the links found for MWE components is another score we explored for ranking our MWE candidates. Here, the accuracy is rather similar giving us 91.7% while using the results of a directional alignment model from Dutch to Spanish. In general, we obtain slightly better results when using word alignment from Dutch to German and Spanish, compared to alignment from Dutch to English.

There emerge several extensions of this work that we wish to address in the future. Alignment types and scoring metrics need to be tested in larger lists of randomly selected MWE candidates to see if the results remain unaltered. We also want to apply some weighting scheme by using the num-

ber of NO_LINKS per expression. Our assumption is that an expression with many NO_LINKS is harder to translate compositionally, and probably an idiomatic or ambiguous expression. Alternatively, an expression with no NO_LINKS is very predictable, thus a literal expression. Finally, another possible improvement is combining several language pairs. There might be cases where idiomatic expressions are conceptualized in a similar way in two languages. For example, a Dutch idiomatic expression with a cognate expression in German might be conceptualized in a different way in Spanish. By combining the entropy or pda scores for NL-EN, NL-DE and NL-ES the accuracy might improve.

Acknowledgments

This research was carried out as part of the research programs for IMIX, financed by NWO and the IRME STEVIN project. We would also like to thank the three anonymous reviewers for their comments on an earlier version of this paper.

References

R.H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania,Philadelphia.

Timothy Baldwin. 2005. Looking for prepositional verbs in corpus data. In *Proc. of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in computational linguistics formalisms and applications*, Colchester, UK.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43th Annual Meeting of the ACL*, pages 597–604, Ann Arbor. University of Michigan.

Miriam Butt. 2003. The light verb jungle. `http://ling.uni-konstanz.de/pages/home/butt/harvard-work.pdf`.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61—74.

Chitra Fernando and Roger Flavell. 1981. *On idiom. Critical views and perspectives*, volume 5 of *Exeter Linguistic Studies*. University of Exeter.

Bart Hollebrandse. 1993. Dutch light verb constructions. Master's thesis, Tilburg University, the Netherlands.

K Imamura, E. Sumita, and Y. Matsumoto. 2003. Automatic construction of machine translation knowledge using translation literalness. In *Proceedings of the 10th EACL*, pages 155–162, Budapest, Hungary.

Adam Kilgarriff and David Tugwell. 2001. Word sketch: Extraction & display of significant collocations for lexicography. In *Proceedings of the 39th ACL & 10th EACL -workshop 'Collocation: Computational Extraction, Analysis and Explotation'*, pages 32–38, Toulouse.

Philipp Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft, available from `http://people.csail.mit.edu/koehn/publications/europarl/`.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

I. Dan Melamed. 1997a. Automatic discovery of non-compositional compounds in parallel data. In *2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, Providence, RI.

I. Dan Melamed. 1997b. Measuring semantic entropy. In *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What and How*, pages 41–46, Washington.

Paola Merlo and Matthias Leybold. 2001. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Procs of the Fifth Computational Natural Language Learning Workshop (CoNLL–2001)*, pages 121–128, Toulouse. France.

Rosamund Moon. 1998. *Fixed expressions and Idioms in English. A corpus-based approach*. Clarendom Press, Oxford.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 20–28, University of Maryland, MD, USA.

Franz Josef Och. 2003. GIZA++: Training of statistical translation models. Available from `http://www.isi.edu/~och/GIZA++.html`.

Ivan Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2001. Multiword expressions: a pain in the neck for NLP. LinGO Working Paper No. 2001-03.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Begoña Villada Moirón. 2005. *Data-driven Identification of fixed expressions and their modifiability*. Ph.D. thesis, University of Groningen.