Multilingual Collocation Extraction: Issues and Solutions

Violeta Seretan
Language Technology Laboratory
University of Geneva
2, rue de Candolle, 1211 Geneva
Violeta.Seretan@latl.unige.ch

Eric Wehrli
Language Technology Laboratory
University of Geneva
2, rue de Candolle, 1211 Geneva
Eric.Wehrli@latl.unige.ch

Abstract

Although traditionally seen as a language-independent task, collocation extraction relies nowadays more and more on the linguistic preprocessing of texts (e.g., lemmatization, POS tagging, chunking or parsing) prior to the application of statistical measures. This paper provides a language-oriented review of the existing extraction work. It points out several language-specific issues related to extraction and proposes a strategy for coping with them. It then describes a hybrid extraction system based on a multilingual parser. Finally, it presents a case-study on the performance of an association measure across a number of languages.

1 Introduction

Collocations are understood in this paper as "idiosyncratic syntagmatic combination of lexical items" (Fontenelle, 1992, 222): *heavy rain*, *light breeze*, *great difficulty*, *grow steadily*, *meet requirement*, *reach consensus*, *pay attention*, *ask a question*. Unlike idioms (*kick the bucket*, *lend a hand*, *pull someone's leg*), their meaning is fairly transparent and easy to decode. Yet, differently from the regular productions, (*big house*, *cultural activity*, *read a book*), collocational expressions are highly idiosyncratic, since the lexical items a headword combines with in order to express a given meaning is contingent upon that word (Mel'čuk, 2003).

This is apparent when comparing a collocation's equivalents across different languages. The English collocation *ask a question* translates as *poser une question* in French (lit., ?*put a question*), and as *fare una domanda*, *hacer una pregunta* in Italian and Spanish (lit., *to make a question*).

As it has been pointed out by many researchers (Cruse, 1986; Benson, 1990; McKeown and Radev, 2000), collocations cannot be described by means of general syntactic and semantic rules. They are arbitrary and unpredictable, and therefore need to be memorized and used as such. They constitute the so-called "semi-finished products" of language (Hausmann, 1985) or the "islands of reliability" (Lewis, 2000) on which the speakers build their utterances.

2 Motivation

The key importance of collocations in text production tasks such as machine translation and natural language generation has been stressed many times. It has been equally shown that collocations are useful in a range of other applications, such as word sense disambiguation (Brown et al., 1991) and parsing (Alshawi and Carter, 1994).

The NLP community fully acknowledged the need for an appropriate treatment of multi-word expressions in general (Sag et al., 2002). Collocations are particularly important because of their prevalence in language, regardless of the domain or genre. According to Jackendoff (1997, 156) and Mel'čuk (1998, 24), collocations constitute the bulk of a language's lexicon.

The last decades have witnessed a considerable development of collocation extraction techniques, that concern both monolingual and (parallel) multilingual corpora.

We can mention here only part of this work: (Berry-Rogghe, 1973; Church et al., 1989; Smadja, 1993; Lin, 1998; Krenn and Evert, 2001) for monolingual extraction, and (Kupiec, 1993; Wu, 1994; Smadja et al., 1996; Kitamura and Mat-

Proceedings of the Workshop on Multilingual Language Resources and Interoperability, pages 40–49,
Sydney, July 2006. ©2006 Association for Computational Linguistics

sumoto, 1996; Melamed, 1997) for bilingual extraction via alignment.

Traditionally, collocation extraction was considered a language-independent task. Since collocations are recurrent, typical lexical combinations, a wide range of statistical methods based on word co-occurrence frequency have been heavily used for detecting them in text corpora. Among the most often used types of lexical association measures (henceforth AMs) we mention: *statistical hypothesis tests* (e.g., binomial, Poisson, Fisher, z-score, chi-squared, t-score, and log-likelihood ratio tests), that measure the significance of the association between two words based on a contingency table listing their joint and marginal frequency, and *Information-theoretic measures* (Mutual Information — henceforth MI — and its variants), that quantity of 'information' shared by two random variables. A detailed review of the statistical methods employed in collocation extraction can be found, for instance, in (Evert, 2004). A comprehensive list of AMs is given (Pecina, 2005).

Very often, in addition to the information on co-occurrence frequency, language-specific information is also integrated in a collocation extraction system (as it will be seen in section 3):

- morphological information, in order to count inflected word forms as instances of the same base form. For instance, *ask questions*, *asks question*, *asked question* are all instances of the same word pair, *ask - question*;

- syntactic information, in order to recognize a word pair even if subject to (complex) syntactic transformations: *ask multiple questions*, *question asked*, *questions that one might ask*.

The language-specific modules thus aim at coping with the problem of morphosyntactic variation, in order to improve the accuracy of frequency information. This becomes truly important especially for free-word order and for high-inflection languages, for which the token(form)-based frequency figures become too skewed due to the high lexical dispersion. Not only the data scattering modify the frequency numbers used by AMs, but it also alters the performance of AMs, if the the probabilities in the contingency table become very low.

Morphosyntactic information has in fact been shown to significantly improve the extraction results (Breidt, 1993; Smadja, 1993; Zajac et al., 2003). Morphological tools such as lemmatizers and POS taggers are being commonly used in extraction systems; they are employed both for dealing with text variation and for validating the candidate pairs: combinations of function words are typically ruled out (Justeson and Katz, 1995), as are the ungrammatical combinations in the systems that make use of parsers (Church and Hanks, 1990; Smadja, 1993; Basili et al., 1994; Lin, 1998; Goldman et al., 2001; Seretan et al., 2004).

Given the motivations for performing a linguistically-informed extraction — which were also put forth, among others, by Church and Hanks (1990, 25), Smadja (1993, 151) and Heid (1994) — and given the recent development of linguistic analysis tools, it seems plausible that the linguistic structure will be more and more taken into account by collocation extraction systems.

The rest of the paper is organized as follows. In section 3 we provide a language-oriented review of the existing collocation extraction work. Then we highlight, in section 4, a series of problems that arise in the transfer of methodology to a new language, and we propose a strategy for dealing with them. Section 5 describes an extraction system, and, finally, section 6 presents a case-study on the collocations extracted for four languages, illustrating the cross-lingual variation in the performance of a particular AM.

3 Overview of Extraction Work

3.1 English

As one might expect, the bulk of the collocation extraction work concerns the English language: (Choueka, 1988; Church et al., 1989; Church and Hanks, 1990; Smadja, 1993; Justeson and Katz, 1995; Kjellmer, 1994; Sinclair, 1995; Lin, 1998), among many others[1].

Choueka's method (1988) detects *n*-grams (adjacent words) only, by simply computing the co-occurrence frequency. Justeson and Katz (1995) apply a POS-filter on the pairs they extract. As in (Kjellmer, 1994), the AM they use is the simple frequency.

Smadja (1993) employs the z-score in conjunction with several heuristics (e.g., the systematic occurrence of two lexical items at the same distance in text) and extracts predicative collocations,

[1] E.g., (Frantzi et al., 2000; Pearce, 2001; Goldman et al., 2001; Zaiu Inkpen and Hirst, 2002; Dias, 2003; Seretan et al., 2004; Pecina, 2005), and the list can be continued.

rigid noun phrases and phrasal templates. He then uses the a parser in order to validate the results. The parsing is shown to lead to an increase in accuracy from 40% to 80%.

(Church et al., 1989) and (Church and Hanks, 1990) use POS information and a parser to extract verb-object pairs, which then they rank according to the mutual information (MI) measure they introduce.

Lin's (1998) is also a hybrid approach that relies on a dependency parser. The candidates extracted are then ranked with MI.

3.2 German

German is the second most investigated language, thanks to the early work of Breidt (1993) and, more recently, to that of Krenn and Evert, such as (Krenn and Evert, 2001; Evert and Krenn, 2001; Evert, 2004) centered on evaluation.

Breidt uses MI and t-score and compares the results accuracy when various parameters vary, such as the window size, presence vs. absence of lemmatization, corpus size, and presence vs. absence of POS and syntactic information. She focuses on N-V pairs[2] and, despite the lack of syntactic analysis tools at the time, by simulating parsing she comes to the conclusion that "Very high precision rates, which are an indispensable requirement for lexical acquisition, can only realistically be envisaged for German with parsed corpora" (Breidt, 1993, 82).

Later, Krenn and Evert (2001) used a German chunker to extract syntactic pairs such as P-N-V. Their work put the basis of formal and systematic methods in collocation extraction evaluation. Zinsmeister and Heid (2003; 2004) focused on N-V and A-N-V combinations identified using a stochastic parser. They applied machine learning techniques in combination to the log-likelihood measure (henceforth LL) for distinguishing trivial compounds from lexicalized ones.

Finally, Wermter and Hahn (2004) identified PP-V combinations using a POS tagger and a chunker. They based their method on a linguistic criterion (that of limited modifiability) and compared their results with those obtained using the t-score and LL tests.

3.3 French

Thanks to the outstanding work of Gross on lexicon-grammar (1984), French is one of the most studied languages in terms of distributional and transformational potential of words. This work has been carried out before the computer era and the advent of corpus linguistics, while automatic extraction was later performed, for instance, in (Lafon, 1984; Daille, 1994; Bourigault, 1992; Goldman et al., 2001).

Daille (1994) aimed at extracting compound nouns, defined a priori by means of certain syntactic patterns, like N-A, N-N, N-à-N, N-de-N, N P Det N. She used a lemmatizer and a POS-tagger before applying a series of AMs, which she then evaluated against a domain-specific terminology dictionary and against a gold-standard manually created from the extraction corpus.

Similarly, Bourigault (1992) extracted noun-phrases from shallow-parsed text, and Goldman et al. (2001) extracted syntactic collocations by using a full parser and applying the LL test.

3.4 Other Languages

In addition to English, German and French, other languages for which notable collocation extraction work was performed, are — as we are aware of — the following:

- Italian: early extraction work was carried out by Calzolari and Bindi (1990) and employed MI. It was followed by (Basili et al., 1994), that made use of parsing information;
- Korean: (Shimohata et al., 1997) used an adjacency *n*-gram model, and (Kim et al., 1999) relied on POS-tagging;
- Chinese: (Huang et al., 2005) used POS information, while (Lu et al., 2004) applied extraction techniques similar to Xtract system (Smadja, 1993);
- Japanese: (Ikehara et al., 1995) was based on an improved *n*-gram method.

As for multilingual extraction via alignment (where collocations are first detected in one language and then matched with their translation in another language), most or the existing work concern the English-French language pair, and the Hansard corpus of Canadian Parliament proceedings. Wu (1994) signals a number of problems

[2] The following abbreviations are used in this paper: N - noun, V - verb, A - adjective, Adv - adverb, Det - determiner, Conj - conjunction, P - preposition.

that non-Indo-European languages pose for the existing alignment methods based on word- and sentence-length: in Chinese, for instance, most of the words are just one or two characters long, and there are no word delimiters. This result suggests that the portability of existing alignment methods to new language pairs is questionable.

We are not concerned here with extraction via alignment. We assume, instead, that multilingual support in collocation extraction means the customization of the extraction procedure for each language. This topic will be addressed in the next sections.

4 Multilingualism: Why and How?

4.1 Some Issues

As the previous section showed, many systems of collocation extraction rely on the linguistic preprocessing of source corpora in order to support the candidate identification process. Language-specific information, such as the one derived from morphological and syntactic analysis, was shown to be highly beneficial for extraction. Moreover, the possibility to apply the association measures on syntactically homogenous material is argued to benefit extraction, as the performance of association measures might vary with the syntactic configurations because of the differences in distribution (Krenn and Evert, 2001).

The lexical distribution is therefore a relevant issue from the perspective of multilingual collocation extraction. Different languages show different proportions of lexical categories (N, V, A, Adv, P, etc.) which are evenly distributed across syntactic types[3]. Depending on the frequency numbers, a given AM could be more suited for a specific syntactic configuration in one language, and less suited for the same configuration in another. Ideally, each language should be assigned a suitable set of AMs to be applied on syntactically-homogenous data.

Another issue that is relevant in the multilingualism perspective is that of the syntactic configurations characterizing collocations. Several such relations (e.g., noun-adjectival modifier, predicate-argument) are likely to remain constant through languages, i.e., to be judged as collocationally interesting in many languages. However, other configurations could be language-specific (like P-N-V in German, whose English equivalent is V-P-N). Yet other configurations might have no counterpart at all in another language (e.g., the French P-A pair *à neuf* is translated into English as a Conj-A pair, *as new*).

Finding all the collocationally-relevant syntactic types for a language is therefore another problem that has to be solved in multilingual extraction. Since a priori defining these types based on intuition does not ensure the necessary coverage, an alternative proposal is to induce them from POS data and dependency relations, as in (Seretan, 2005).

The morphoyntactic differences between languages also have to be taken into account. With English as the most investigated language, several hypotheses were put forth in extraction and became common place.

For instance, using a 5-words window as search space for collocation pairs is a usual practice, since this span length was shown sufficient to cover a high percentage of syntactic co-occurrences in English. But — as suggested by other researchers, e.g., (Goldman et al., 2001) —, this assumption does not necessary hold for other languages.

Similarly, the higher inflection and the higher transformation potential shown by some languages pose additional problems in extraction, which were rather ignored for English. As Kim et al. (1999) notice, collocation extraction is particularly difficult in free-order languages like Korean, where arguments scramble freely. Breidt (1993) also pointed out a couple of problems that makes extraction for German more difficult than for English: the strong inflection for verbs, the variable word-order, and the positional ambiguity of the arguments. She shows that even distinguishing subjects from objects is very difficult without parsing.

4.2 A Strategy for Multilingual Extraction

Summing up the previous discussion, the customization of collocation extraction for a given language needs to take into account:

- the syntactic configurations characterizing collocations,
- the lexical distribution over syntactic configurations,
- the adequacy of AMs to these configurations.

[3] For instance, V-P pairs are more represented in English than in other languages (as phrasal verbs or verb-particle constructions).

These are language-specific parameters which need to be set in a successful multilingual extraction procedure. Truly multilingual systems have not been developed yet, but we suggest the following strategy for building such a system:

A. parse the source corpus, extract all the syntactic pairs (e.g., head-modifier, predicate-argument) and rank them with a given AM,

B. analyze the results and find the syntactic configurations characterizing collocations,

C. evaluate the adequacy of AMs for ranking collocations in each syntactic configuration, and find the most convenient mapping configurations - AMs.

Once customized for a language, the extraction procedure involves:

Stage 1. parsing the source corpus for extracting the lexical pairs in the relevant, language-specific syntactic configurations found in step B;

Stage 2. ranking the pairs from each syntactic class with the AM assigned in step C.

5 A Multilingual Collocation Extractor Based on Parsing

Ever since the collocation was brought to the attention of linguists in the framework of contextualism (Firth, 1957; Firth, 1968), it has been preponderantly seen as a pure statistical phenomenon of lexical association. In fact, according to a well-known definition, "a collocation is an arbitrary and recurrent word combination" (Benson, 1990).

This approach was at the basis of the computational work on collocation, although there exist an alternative approach — the linguistic, or lexicographic one — that imposes a restricted view on collocation, which is seen first of all as an expression of language.

The existing extraction work (section 3) shows that there is a growing interest in adopting the more restricted (linguistic) view. As mentioned in section 3, the importance of parsing for extraction was confirmed by several evaluation experiments. With the recent development in the field of linguistic analysis, hybrid extraction systems (i.e., systems relying on syntactical analysis for collocation extraction) are likely to become the rule rather than the exception.

Our system (Goldman et al., 2001; Seretan and Wehrli, 2006) is — to our knowledge — the first to perform the full syntactic analysis as support for collocation extraction; similar approaches rely on dependency parsers or on chunking.

It is based on a symbolic parser that was developed over the last decade (Wehrli, 2004) and achieves a high level of performance, in terms of accuracy, speed and robustness. The languages it supports are, for the time being, French, English, Italian, Spanish and German. A few other languages are being also implemented in the framework of a multilingualism project.

Provided that collocation extraction can be seen as a two-stage process (where, in stage 1, collocation candidates are identified in the text corpora, and in stage 2, they are ranked according to a given AM, cf. section 4.2), the role of the parser is to support the first stage. A pair of lexical items is selected as a candidate only if there exist a syntactic relation holding between the two items.

Unlike the traditional, window-based methods, candidate selection is based on syntactic proximity (as opposed to textual proximity). Another peculiarity of our system is that candidate pairs are identified as the parsing goes on; in other approaches, they are extracted by post-processing the output of syntactic tools.

The candidate pairs identified are classified into syntactically homogenous sets, according to the syntactic relations holding between the two items. Only certain predefined syntactic relations are kept, that were judged as collocationally relevant after multiple experiments of extraction and data analysis (e.g., adjective-noun, verb-object, subject-verb, noun-noun, verb-preposition-noun). The sets obtained are then ranked using the log-likelihood ratios test (Dunning, 1993).

More details about the system and its performance can be found in (Seretan and Wehrli, 2006). The following examples (taken from the extraction experiment we will describe below) illustrate its potential to detect collocation candidates, even if these are subject to complex syntactic transformations:

1.a) *atteindre objectif* (Fr): Les *objectifs* fixés à l'échelle internationale visant à réduire les émissions ne peuvent pas être *atteints* à l'aide de ces seuls programmes.

1.b) *accogliere emendamento* (It):

Posso pertanto *accogliere* in parte e in linea di principio gli *emendamenti* nn. 43-46 e l'emendamento n. 85.

1.c) *reforzar cooperación* (Es): Queremos permitir a los pases que lo deseen *reforzar*, en un contexto unitario, su *cooperación* en cierto número de sectores.

The collocation extractor is part of a bigger system (Seretan et al., 2004) that integrates a concordancer and a sentence aligner, and that supports the visualization, the manual validation and the management of a multilingual terminology database. The validated collocations are used for populating the lexicon of the parser and that of a translation system (Wehrli, 2003).

6 A Cross-Lingual Extraction Experiment

A collocation extraction experiment concerning four different languages (English, Spanish, French, Italian) has been conducted on a parallel subcorpus of 42 files from the European Parliament proceedings. Several statistics and extraction results are reported in Table 1.

Statistics	**English**	**Spanish**	**Italian**	**French**
tokens	2526403	2666764	2575858	2938118
sent/file	2329.1	2513.7	2331.6	2392.8
complete parses	63.4%	35.5%	46.8%	63.7%
tokens/sent	25.8	25.3	26.3	29.2
extr. pairs (tokens)	617353	568998	666122	565287
token/type	2.6	2.5	2.3	2.3
LL is def.	85.9%	90.6%	83.5%	92.8%

Table 1: Extraction statistics

We computed the distribution of pair tokens according to the syntactic type and noted that the most marked distributional difference among these languages concern the following types: N-A (7.12), A-N (4.26), V-O (2.68), V-P (4.16), N-P-N (3.81)[4].

Unsurprisingly, the Romance languages are less different in terms of syntactic co-occurrence distribution, and the deviation of English from the Romance mean is more pronounced — in particular, for N-A (9.72), V-P (5.63), A-N (5.25), N-P-N (4.77), and V-O (3.57). These distributional differences might account for the types of collocations highlighted by a particular AM (such as LL) in a language vs. another. Figure 1 displays the relative proportions of 3 syntactic types — adjective-noun, subject-verb and verb-object — that can be found at different levels in the significance list returned by LL.

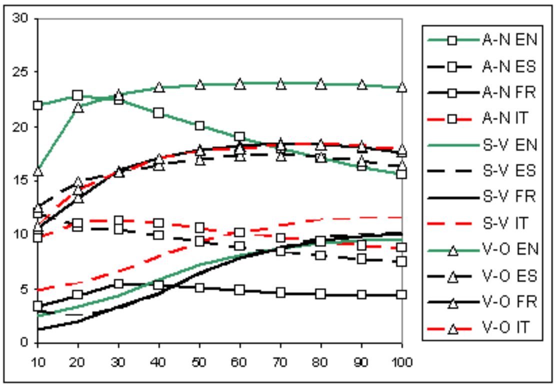

Figure 1: Cross-lingual proportions of A-N, S-V and V-O pairs at different levels in the significance lists

We performed a contrastive analysis of results, by carrying out a case-study aimed at checking the LL performance variability across languages. The study concerned the verb-object collocations having the noun *policy* as the direct object. We specifically focused on the best-scored collocation extracted from the French corpus, namely *mener une politique* (lit., *conduct a policy*).

We looked at the translation equivalents of its 74 instances identified by our extraction system in the corpus. The analysis revealed that — at least in this particular case — the verbal collocates of this noun are highly scattered: *pursue, implement, conduct, adopt, apply, develop, have, draft, launch, run, carry out* for English; *practicar, llevar a cabo, desarrollar, realizar, aplicar, seguir, hacer, adoptar, ejercer* for Spanish; *condurre, attuare, portare avanti, perseguire, praticare, adottare, fare* for Italian (among several others). Some of the collocates (those listed first) are more prominently used. But generally they are highly dispersed, and this might indicate a bigger difficulty for LL to pinpoint the best collocate in a language vs. another.

We also observed that quite frequently (in about 25% of the cases) the collocation did not conserve its syntactic configuration. Either the verb — here,

[4] The numbers represent the values the standard deviation of the relative percentages in the whole lists of pairs.

the equivalent for the French *mener* — is omitted in translations (like in 2.b below):

2.a) des contradictions existent dans la politique qui est menée (Fr);

2.b) we are dealing with contradictory policy (En),

or, in a few other cases, the whole collocation disappears, since paraphrased with a completely different syntactic construction:

3.a) direction qui a mené une politique insensée de réduction de personnel (Fr);

3.b) a management that foolishly engaged in staff reductions (En).

In order to quantify the impact such factors have on the performance of the AM considered, we further scrutinized the collocates list for *politique* proposed by LL test for each language (see Table 2). The rank of a pair in the whole list of verb-object collocations extracted, as assigned by the LL test, is shown in the last column. In these significance lists, the collocations with *politique* as an object constitute a small fraction, and from these, only the top collocations are displayed in Table 2. The threshold was manually defined in accordance with our intuition that the lower-scored pairs observed manifest less a collocational strength. It happens to be situated around the LL value of 20 for each language (and is of course specific to the size of our corpus and to the number of V-O tokens identified therein).

If we consider the LL rank as the success measure for collocate detection, we can infer that the collocates of the word under investigation are easier to found in French, as compared to English, Italian or Spanish, because the value in the first row of the last column is smaller. This holds if we are interested in only one (the most salient) collocate for a word.

If we measure the success of retrieving all the collocates (by considering, for instance, the speed to access them in the results list — the higher the rank, the better), then French can be again considered the easiest because overall, the positions in the V-O list are higher (i.e., the mean of the rank column is smaller) with respect to Spanish, Italian and, respectively, English.

This latter result corresponds, approximately, to the order given by relative proportion of V-O pairs in each language (Spanish 15.12%, French 15.14%, Italian 17.06%, and English 20.82%). Given that in English V-O pairs are more numerous and the verbs also participate in V-P constructions, it might seem reasonable to expect lower LL scores for V-O collocations in English vs. the other 3 languages.

Language	**collocate**	**freq**	**LL score**	**rank**
French	mener	74	376.8	45
politique	élaborer	17	50.1	734
	adapter	5	48.3	780
	axer	8	41.4	955
	pratiquer	9	39.7	1011
	développer	13	28.1	1599
	adapter	8	25.2	1867
	poursuivre	11	24.4	1943
English	pursue	39	214.9	122
policy	implement	38	108.7	325
	develop	30	81.1	473
	conduct	8	28.9	2014
	harmonize	9	28.2	2090
	gear	5	27.7	2201
	need	25	24.9	2615
	apply	16	23.3	2930
Spanish	practicar	17	98.7	246
política	desarrollar	27	82.4	312
	aplicar	25	65.7	431
	seguir	17	33.5	1003
	coordinar	8	31.0	1112
	basar	11	25.1	1473
	orientar	6	22.5	1707
	adaptar	5	20.0	1987
	construir	6	19.4	2057
Italian	attuare	23	79.5	382
politica	perseguire	14	46.4	735
	praticare	8	37.6	976
	seguire	18	30.2	1314
	portare	12	29.7	1348
	rivedere	9	26.0	1607
	riformare	7	25.6	1639
	sviluppare	12	22.1	1975
	adottare	20	21.2	2087

Table 2: Verbal collocates for the headword *policy*

In general, we expect a correlation between extraction difficulty and the distributional properties of co-occurrence types.

7 Conclusion

The paper pointed out several issues that occur in transfering a hybrid collocation extraction methodology (that combines linguistic with statistic information) to a new language.

Besides the questionable availability of language-specific text analysis tools for the new language, a number of issues that are relevant to extraction proper were addressed: the changes in the distribution of (syntactic) word pairs, and the need to find, for each language, the most appropriate association measure to apply for each syntactic type (given that AMs are sensitive to distributions and syntactic types); the lack of a priori defined syntactic types for a language; and, finally, the portability of some widely used techniques (such as the window method) from English to other languages exhibiting a higher word order freedom.

It is again in the multilingualism perspective that the inescapable need for preprocessing the text emerged (cf. different researchers cited in section 3): highly inflected languages need lemmatizers, free-word order languages need structural information in order to guarantee acceptable results. As language tools become nowadays more and more available, we expect the collocation extraction (and terminology acquisition in general) to be exclusively performed in the future by relying on linguistic analysis. We therefore believe that multilingualism is a true concern for collocation extraction.

The paper reviewed the extraction work in a language-oriented fashion, while mentioning the type of linguistic preprocessing performed whenever it was the case, as well as the language-specific issues identified by the authors. It then proposed a strategy for implementing a multilingual extraction procedure that takes into account the language-specific issues identified.

An extraction system for four different languages, based on full parsing, was then described. Finally, an experiment was carried out as a case study, which pointed out several factors that might determine a particular AM to perform differently across languages. The experiment suggested that log-likelihood ratios test might highlight certain verb-object collocations easier in French than in Spanish, Italian and English (in terms of salience in the significance list).

Future work needs to extend the type of cross-linguistic analysis initiated here, in order to provide more insights on the differences expected at extraction between one language and another and on the responsible factors, and, accordingly, to defines strategies to deal with them.

Acknowledgements

The research described in this paper has been supported in part by a grant from the Swiss National Foundation (No. 101412-103999).

References

Hiyan Alshawi and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. 1994. A ”not-so-shallow” parser for collocational analysis. In *Proceedings of the 15th conference on Computational linguistics*, pages 447–453, Kyoto, Japan. Association for Computational Linguistics.

Morton Benson. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–35.

Godelieve L. M. Berry-Rogghe. 1973. The computation of collocations and their relevance to lexical studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*, pages 103–112. Edinburgh.

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 977–981, Nantes, France.

Elisabeth Breidt. 1993. Extraction of V-N-collocations from text corpora: A feasibility study for German. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, U.S.A.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, pages 264–270, Berkeley, California.

Nicoletta Calzolari and Remo Bindi. 1990. Acquisition of lexical information from a large textual Italian corpus. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 54–59, Helsinki, Finland.

Yaacov Choueka. 1988. Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In *Proceedings of the International Conference on User-Oriented Content-Based Text and Image Handling*, pages 609–623, Cambridge, U.S.A.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1989. Parsing, word associations and typical predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*, pages 103–112, Pittsburgh. Carnegie Mellon University.

D. Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Gaël Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, pages 41–48, Sapporo, Japan.

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

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

John Rupert Firth, 1957. *Papers in Linguistics 1934-1951*, chapter Modes of Meaning, pages 190–215. Oxford Univ. Press, Oxford.

J. R. Firth. 1968. A synopsis of linguistic theory, 1930–55. In F.R. Palmer, editor, *Selected papers of J. R. Firth, 1952-1959*. Indiana University Press, Bloomington.

Thierry Fontenelle. 1992. Collocation acquisition from a corpus or from a dictionary: a comparison. *Proceedings I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere*, pages 221–228.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 2(3):115–130.

Jean-Philippe Goldman, Luka Nerima, and Eric Wehrli. 2001. Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocations*, pages 61–66, Toulouse, France.

Maurice Gross. 1984. Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 22nd conference on Association for Computational Linguistics*, pages 275–282, Morristown, NJ, USA.

Franz Iosef Hausmann. 1985. Kollokationen im deutschen wörterbuch. ein beitrag zur theorie des lexikographischen beispiels". In Henning Bergenholtz and Joachim Mugdan, editors, *Lezikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch.*, Lexicographica. Series Major 3, pages 118–129.

Ulrich Heid. 1994. On ways words work together - research topics in lexical combinatorics. In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, and P. Vossen, editors, *Proceedings of the VIth Euralex International Congress (EURALEX '94)*, pages 226–257, Amsterdam.

Chu-Ren Huang, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 48–55, Jeju Island, Republic of Korea.

Satoru Ikehara, Satoshi Shirai, and Tsukasa Kawaoka. 1995. Automatic extraction of uninterrupted collocations by n-gram statistics. In *Proceedings of first Annual Meeting of the Association for Natural Language Processing*, pages 313–316.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistis properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Seonho Kim, Zooil Yang, Mansuk Song, and Jung-Ho Ahn. 1999. Retrieving collocations from Korean text. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 71–81, Maryland, U.S.A.

Mihoko Kitamura and Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 79–87, Copenhagen, Denmark, August.

Göran Kjellmer. 1994. *A Dictionary of English Collocations*. Claredon Press, Oxford.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, U.S.A.

P. Lafon. 1984. *Dépouillement et statistique en léxicometrie*. Slatkine-Champion, Paris.

Michael Lewis. 2000. *Teaching Collocations. Further Developments In The Lexical Approach*. Language Teaching Publications, Hove.

Dekang Lin. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, pages 57–63, Montreal.

Qin Lu, Yin Li, and Ruifeng Xu. 2004. Improving Xtract for Chinese collocation extraction. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 333–338.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*, pages 507–523. Marcel Dekker, New York, U.S.A.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL'97)*, pages 305–312, Madrid, Spain.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Claredon Press, Oxford.

Igor Mel'čuk. 2003. Collocations: définition, rôle et utilité. In Francis Grossmann and Agnès Tutin, editors, *Les collocations: analyse et traitement*, pages 23–32. Editions "De Werelt", Amsterdam.

Darren Pearce. 2001. Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46, Pittsburgh, U.S.A.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan, June.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City.

Violeta Seretan and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of COLING/ACL 2006*. To appear.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2004. A tool for multi-word collocation extraction and visualization in multilingual corpora. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, pages 755–766, Lorient, France.

Violeta Seretan. 2005. Induction of syntactic collocation patterns from generic syntactic relations. In *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1698–1699, Edinburgh, Scotland, July.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. 1997. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 476–481, Madrid, Spain.

John Sinclair. 1995. *Collins Cobuild English Dictionary*. Harper Collins, London.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Eric Wehrli. 2003. Translation of words in context. In *Proceedings of Machine Translation Summit IX*, pages 502–504, New Orleans, Lousiana, U.S.A.

Eric Wehrli. 2004. Un modèle multilingue d'analyse syntaxique. In A. Auchlin et al., editor, *Structures et discours - Mélanges offerts à Eddy Roulet*, pages 311–329. Éditions Nota bene, Québec.

Joachim Wermter and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 980–986, Geneva, Switzerland.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, pages 80–87, Las Cruces (New Mexico), U.S.A.

Diana Zaiu Inkpen and Graeme Hirst. 2002. Acquiring collocations for lexical choice between near-synonyms. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 67–76, Philadephia, Pennsylvania.

Rémi Zajac, Elke Lange, and Jin Yang. 2003. Customizing complex lexical entries for high-quality MT. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, U.S.A.

Heike Zinsmeister and Ulrich Heid. 2003. Significant triples: Adjective+Noun+Verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003), Budapest*.

Heike Zinsmeister and Ulrich Heid. 2004. Collocations of complex nouns: Evidence for lexicalisation. In *Proceedings of KONVENS 2004*, Vienna, Austria.