# How to Compile a Bilingual Collocational Lexicon Automatically

**Frank Smadja, smadja@cs.columbia.edu**

**Columbia University**
**Computer Science Department**
**New York, NY 10027**

## 1.0 Introduction and Motivation[1]

Consider the following two sentences:

*(1e) "The Government is wasting millions of dollars by sending monthly pension cheques to wealthy **senior citizens** and **baby bonuses** to families who do not need the money."*

*(1f) "Le gouvernement gaspille des millions de dollars en envoyant des chèques de pension tous les mois aux **personnes âgées** riches et des **allocations** aux familles qui n'en ont pas besoin."*

These sentences are extracted from a corpus of the proceedings of the Canadian Parliament, also called the Hansards corpus. As required by law, the Hansards corpus have both the English and the French for each sentence.The corpus consists of a number of pairs of files, one written in English and the other one in French. We used a version of the Hansards in which the sentences have been aligned with their translations as described in [Church91][2]. Sentence (1f) is thus the translation in French of Sentence (1e).[3]

Automatically producing (1e) from (1f) involves many issues that have yet to be solved. In this paper we address the simpler task of finding correct translations for collocations, in order to produce bilingual lexical information. More precisely, in the above sentences the translation for *"senior citizens"* is *"personnes âgées"* and the translation for *"baby bonuses"* is *"allocations."* This actually raises several problems:

1. The translation of most collocations is not predictable in terms of the meaning of the individual words, and/or the meaning of the whole collocation. Although the usual translation for *"citizen"* is *"citoyen,"* the latter is not used in the translation for "*senior citizen.*" Without the knowledge of the proper collocations, one would attempt a word-by-word translation and end up with incorrect translations (*e.g., "un bonus pour bébés"*).
2. As discussed in [Smadja 92], most dictionaries do not include collocations. This is true both for monolingual dictionaries and bilingual ones.
3. Some collocations translate into single words. More generally, a collocation of n words ($1 <= n$) might

[1] This research is supported ONR grant N00014-89-J-1782. Proceedings of the AAAI workshop on statistically based NLP techniques, July, 1992.

[2] We would like to thank Ken Church and the Bell Laboratories here for providing us with the aligned corpus.

[3] Although in actuality, (1e) might have been translated from (1f).

translate in a collocation of p words *(1 <= p)*, in which n and p are different.

Problem 1 is the motivation for our work. It indicates that a collocational bilingual dictionary is needed to do proper translation. Furthermore, since currently available dictionaries are not adequate there is a need for creating such collocational dictionaries.

In this paper, we propose a technique for constructing bilingual collocation dictionaries completely automatically. The technique we propose first identifies a set of collocations in one language and then attempts to translate them using the Hansards as training data. To do this, we propose to use Xtract, a collocation compiler [Smadja 92], to identify collocations and to use mutual information statistics to translate the collocations into the other language. The algorithm we describe is an iterative method that builds the translation of a given collocation by adding words one by one. This technique allows a collocation containing n words to be translated into a collocation of p words. The paper describes the proposed algorithm and shows how it is applied in the translation of the following three collocations: "*senior citizen.*," "*Madam Speaker,*" and "*election campaign.*"

## 2.0 Related work

The work we describe in this paper can be viewed as an extension of the work done by [Gale & Church 91] and [Brown *et.al* .91a] on bilingual sentence alignment. Both works use purely statistical techniques to identify sentence pairing in corpora similar to the Hansards. Although aligning sentences might seem like a relatively minor task, it is very productive as it provides a starting point from which to pursue research. As mentioned before, in the work we present here, we used an aligned corpus as input data.

Another thread of research attempts to do translation using statistical techniques only. [Brown *et.al.*91b] use a stochastic language model based on the techniques used in speech recognition [Bahl *et.al.*83], combined with translation probabilities compiled on the aligned corpus in order to do sentence translation. Although the project is still at an early stage, it already produces quality translation for simple sentences without any linguistic or semantic information. In the process of translating sentences, they also align groups of words to other groups of words. However, this is only a substep in the translation process and it is not made to produce any sort of lexicon. Moreover, it is not clear how these alignments could be used in non-statistical approaches. In contrast, we use statistical techniques in order to provide bilingual lexical information that could be used across a variety of applications.

## 3.0 Aligning words using mutual informations scores.

The first stage in aligning collocations is not novel (*i.e.*,[Gale & Church 91] and [Brown *et.al.*91b]). It consists of using mutual information scores to evaluate the correlation of pairs of French and English words. These scores will then be used in the next stages to select candidates for inclusion in collocations.

The mutual information between two events is usually defined as:

$$\langle 1\rangle \mu (e,f) = \log \left(\frac{p(e \wedge f)}{p(e) \times p(f)}\right)$$

where *e* and *f* are two separate events, and *p(x)* denotes the probability of appearance of x. $\mu(e,f)$ measures how the two events are correlated.

We applied Equation (1) as follows. Let $S'_n$ be a given sentence chosen randomly in the corpus, let $S|_n^E$ be the English version of this sentence, and $S|_n^F$ be the French alignment. Let *E* and *F* be respectively an English word and a French word, and let *e* and *f* respectively be the events that $F \in S|_n^F$ and that $E \in S|_n^E$. Using the corpus as training data, we can compute probabilities using a simple maximum likelihood method, and thus the mutual information of the two events can be computed as follows:

$$\langle 2\rangle \mu (e,f) = \log \left(\frac{|S(E) \cap S(F)|}{|S(E)| \cdot |S(F)|} \cdot N\right)$$

where S(W) denotes the set of sentences containing the word *W*, and *N* denotes the total number of sentences in the training corpus.

Using Equation 2 on the Hansards corpus, we compiled mutual information scores for most pairs of possible English and French words and we kept all pairs with mutual information scores significantly greater than 0. Table 1 shows a randomly selected subset of these aligned words which has been sorted in decreasing score for this paper. Most words listed in the Table are actual translation of one another, for example, the days of the week, the months, and mostly unambiguous words such as *afternoon* and *après midi*, *mandatory* and *obligatoire* have been correctly associated. Such data could already be used efficiently by humans. However, we notice that there are some discrepancies, due to several factors. We have identified the following two factors as accounting for a vast majority of the incorrectly aligned words:

- Ambiguous words.
  Ambiguous words are often translated into several words depending on the context. In the table, we see that, for example, *inflation* is aligned with the French *inflation*. This is only true when *inflation* means "*an increase in the volume of money and credit,*" but the translation is not correct when *inflation* means "*the act of inflating*." The more appropriate translation would be "*gonflage*" or "*gonflement.*" Although the two senses are related, French has two different and unrelated words.
- Collocations.
  Some words are used as part of a collocation, and a collocation often translates into another collocation. As explained before, collocations do not translate well across languages. So that simple mutual information scores might provide wrong associations in which one word of a given collocation is associated with any word of the corresponding collocation. In the table, examples of wrong associations due to collocations are in bold fonts. For example, the proper translation for "*senior citizen*" is "*personnes agées*," and the correct translation for "*election campaign*," is "*campagne electorale*." In the table, we see that *campaign* gets associated either with *electorale (*which means "*relating to an election*"), or with "*campagne*" (which is only true in the context of this collocation), and *senior* gets associated with *agées* (which means *old*).

**TABLE 1.** Some word alignments

| English | French | Score |
|---|---|---|
| october | octobre | 2.766482 |
| inflation | inflation | 2.760550 |
| friday | vendredi | 2.756804 |
| december | décembre | 2.743461 |
| **scotia** | **nouvelle-écosse** | **2.688104** |
| bay | bay | 2.676008 |
| afternoon | aprés-midi | 2.669028 |
| **nova** | **nouvelle-écosse** | **2.649809** |
| war | guerre | 2.642909 |
| patent | brevets | 2.605734 |
| madam | madame | 2.590855 |
| thousand | milliers | 2.586640 |
| mine | mines | 2.579116 |
| solicitor | solliciteur | 2.561797 |
| native | autochtones | 2.544999 |
| mandatory | obligatoire | 2.526343 |
| morning | matin | 2.525235 |
| **expansion** | **regionale** | 2.520080 |
| **campaign** | **campagne** | **2.515319** |
| debt | dette | 2.514616 |
| welfare | bien-être | 2.505738 |
| hill | colline | 2.503827 |
| progress | progrès | 2.496735 |
| **campaign** | **electorale** | **2.492681** |
| expansion | expansion | 2.492560 |
| **supervision** | **obligatoire** | **2.492257** |
| **madam** | **présidente** | **2.488770** |
| mandate | mandat | 2.481933 |
| supervision | surveillance | 2.473921 |
| withdraw | retirer | 2.469366 |
| men | hommes | 2.469220 |
| **constituent** | **électeurs** | **2.467424** |
| gas | gaz | 2.457627 |
| **expansion** | **industrielle** | **2.456365** |
| water | eau | 2.454382 |
| defend | défendre | 2.443816 |
| growth | croissance | 2.438468 |
| cabinet | cabinet | 2.438364 |
| **senior** | **agées** | **2.426776** |
| children | enfants | 2.422183 |

In this paper, we do not address the case of ambiguous words, but we are mostly concerned with the case of collocations.

## 4.0 Using Xtract for finding English collocations

Providing translation for collocations presupposes that collocations are already know in one language L1 and that one wants to express them in an other language L2. To identify collocations, we propose to use a collocation compiler, Xtract [Smadja 92]. Using Xtract allows us to start the translation process by providing us with a set of collocations to translate. It thus greatly reduces the search space in identifying many-to-many associations between English and French. A year of the Hansards has a vocabulary of more than 20,000 words so that the search space for collocations of length 2 to 8 would be in the order of $10^{33}$ whereas Xtract produces only several thousand collocation of length 2 to 8.

### 4.1 Xtract, an Overview

Described in [Smadja & McKeown 90, Smadja 92] Xtract is a tool for compiling collocations from an unstructured free text corpus. Xtract produces a wide range of collocations. In particular, Xtract produces flexible collocations of the type *"to make a decision,"* in which the words can be inflected, the word order might change and the number of additional words vary with the examples. In [Smadja 92] we show that Xtract can identify such collocations with a precision of 80%. Xtract also produces compounds, such as *"The Dow Jones average of 30 industrial stock,"* which are non flexible collocations. In this paper we only use the collocations of type compounds which have also been identified by other techniques such as [Chòueka *et.al.* 83].

### 4.2 Compiling Collocations with Xtract

We have used Xtract on the English version of the Hansards to compile compound collocations. Among the collocations retrieved are: *"Madam Speaker," "the election campaign," "regional industrial expansion," "the Prime Minister,"* and *"senior citizen."* In this paper we are looking for the translation of *"Madam Speaker," "the election campaign,"* and *"senior citizen."*

**TABLE 2.** Possible translations of senior

| | | |
|---|---|---|
| senior | agées | 2.426776 |
| senior | troisième | 1.200094 |
| senior | direction | 1.226118 |
| senior | fonctionnaires | 1.238787 |
| senior | citoyens | 1.156518 |
| senior | sénat | 0.915873 |
| senior | américain | 0.866879 |
| senior | population | 0.813242 |
| senior | niveau | 0.929738 |
| senior | circonscription | 0.890250 |
| senior | femmes | 0.706010 |
| senior | sécurité | 0.703194 |
| senior | jeunes | 0.650396 |
| senior | postes | 0.636763 |
| senior | mois | 0.615607 |
| senior | comment | 0.604284 |
| senior | service | 0.587988 |
| senior | ministère | 0.641287 |
| senior | conservateur | 0.713594 |
| senior | aurait | 0.536152 |
| senior | décision | 0.534065 |
| senior | contre | 0.609833 |
| senior | nombre | 0.464127 |
| senior | prendre | 0.531391 |
| senior | finances | 0.494123 |
| senior | parlementaire | 0.452941 |
| senior | conseil | 0.487362 |
| senior | santé | 0.521901 |
| senior | personnes | 1.621361 |
| senior | services | 0.656421 |
| senior | gens | 0.460902 |
| senior | mesures | 0.460181 |
| senior | programme | 0.437212 |
| senior | encore | 0.292530 |
| senior | société | 0.278405 |

## 5.0 Translating Collocations

### 5.1 Hypotheses and Overall Description.

The technique we propose uses compound collocations as identified by Xtract as seeds, and attempts to provide translation for them. Theoretically this process must be applied both from French to English and from English to French, so that many-to-one and one-to-many grouping could be identified. However, we only applied it from English to French for the moment. The assumption on which the translation process is based on says that if two collocations, E and F, are translations of one another, then all the words in E are correlated with all the words in F. The algorithm we propose to use attempts to build the translation of a seed English collocation by incrementally adding single words to its French translation.

**TABLE 3.** Possible translations of citizen

| | | |
|---|---|---|
| citizen | citoyens | 2.090491 |
| citizen | âgées | 2.069875 |
| citizen | personnes | 1.293101 |
| citizen | pétition | 1.271266 |
| citizen | ottawa | 0.997750 |
| citizen | habitants | 0.993420 |
| citizen | troisième | 0.981948 |
| citizen | lois | 0.922442 |
| citizen | bien-être | 0.922442 |
| citizen | honneur | 0.875496 |
| citizen | circonscription | 0.824966 |
| citizen | qualité | 0.823696 |
| citizen | groupe | 0.806067 |
| citizen | présenter | 0.783948 |
| citizen | sécurité | 0.783605 |
| citizen | protéger | 0.780380 |
| citizen | matin | 0.766204 |
| citizen | décidé | 0.765796 |
| citizen | population | 0.736677 |
| citizen | justice | 0.735672 |
| citizen | accès | 0.711306 |
| citizen | canadiens | 0.681060 |
| citizen | eux | 0.678879 |
| citizen | plupart | 0.677068 |
| citizen | santé | 0.675888 |
| citizen | droit | 0.663617 |
| citizen | moyen | 0.659368 |
| citizen | services | 0.659038 |
| citizen | vie | 0.647867 |
| citizen | payer | 0.646507 |
| citizen | groupes | 0.636491 |
| citizen | aider | 0.633778 |
| citizen | parlement | 0.629732 |
| citizen | besoin | 0.619423 |

**TABLE 4.** Intermediate translations

| | | | |
|---|---|---|---|
| senior | citizen | âgées | 2.638035, |
| senior | citizen | personnes | 1.831326, |
| senior | citizen | troisième | 1.476078, |
| senior | citizen | citoyens | 1.385506, |
| senior | citizen | population | 1.083623, |
| senior | citizen | circonscription | 1.073481, |
| senior | citizen | sécurité | 0.969971, |
| senior | citizen | niveau | 0.878886, |
| senior | citizen | services | 0.867681, |
| senior | citizen | jeunes | 0.820263, |
| senior | citizen | conservateur | 0.817643, |
| senior | citizen | contre | 0.783304, |
| senior | citizen | santé | 0.757519, |
| senior | citizen | nombre | 0.730904, |
| senior | citizen | parlementaire | 0.719718, |
| senior | citizen | gens | 0.696520, |
| senior | citizen | femmes | 0.671757, |
| senior | citizen | américain | 0.663165, |
| senior | citizen | aurait | 0.581080, |
| senior | citizen | encore | 0.513549, |
| senior | citizen | programme | 0.377404, |
| senior | citizen | service | 0.331886, |
| senior | citizen | prendre | 0.327677, |
| senior | citizen | mesures | 0.321192, |
| senior | citizen | société | 0.285544, |
| senior | citizen | décision | 0.240174, |

## 5.2 The algorithm

The algorithm is an iterative algorithm that constructs the translation for a given collocation on a word by word basis. Let $\{e'_1,\ldots,e'_n\}$ be an English collocation as identified by Xtract. The algorithm is as follows:

1. **Compute** $\bigcap_{j} f|_{j}^{i} = S'_1$ In which $\mu(e,f) > 1$ and $S'_i$ is the set of French words that are correlated with each of the words of $\{e'_1,\ldots,e'_n\}$.
   Let i = 1.
2. **Sort** the elements of $S'_i$ by decreasing mutual information scores.
3. For each subset of size i+1 of $S'_i$, **compute** the mutual information of all its elements taken as separate events.
4. **Remove** all the sets containing non correlated elements.

5. **If** there is no remaining subset of size i+1, **then** produce the subset of size i with the highest mutual information score with the seed English collocation and **go to** Step 6.
   **Otherwise, Increment i and go to** Step 2.
6. **End.**

In the above algorithm, we define the mutual information of a set of size p > 1 an event x as the mutual information of conjunction of the pth element of the set and the remaining subset of size (p -1), with the event x.

### 5.3 Some Preliminary Results

We have experimented with the above algorithm for three English collocations, and we have reached the correct French equivalent in all three cases. Although this does not allow us to determine the validity of the algorithm we consider it an encouraging result. In the rest of this section we show the algorithm for the collocation: *"senior citizen."*

Table 2 and 3, list the associations of the words *senior* and *citizen* respectively. As can be seen from Tables 2 and 3, *senior* has some 30 possible translations and *citizen* has some 130. Applying Step 1 of the algorithm we compute $S'_i$, which consists of the following 26 words: *troisième, societé, services, service, sécurité, santé, programme, prendre, population, personnes, parlementaire, nombre, niveau, mesures, jeunes, gens, femmes, encore, décision, contre, conservateur, citoyens, circonscription, aurait, américain, àgées.*

Applying Step 2 of the algorithm, we compute the mutual information of each of the above words with the seed collocation. Table 4, indicates these results.

After the application of Step 3, only one subset of size 2 remained: *"personnes àgées"* which is the correct translation for "*senior citizen.*" Which then terminates the algorithm.

The application of the same algorithm on *"election campaign"* and *"Madam Speaker,"* also produced the correct results: *"campagne electorale"* and *"Madame la Présidente,"* in the same number of steps. The translation of *"Madam Speaker"* is obviously specific to this corpus and cannot be generalized. In contrast, the translation of *"election campaign"* is general and valid across domains. In addition, it is interesting because it is a problem to translate noun-noun compounds in French since French syntax does not allow for such constructs.

We are currently working on testing this algorithm for more complex cases, i.e., cases in which an English collocation of size n is translated in a French collocation of a different size. In particular when the French translation consists of a single word. We are also evaluating the use of statistics other than mutual information that would bring better results. In a next stage, we will apply the technique to a large number of collocations and we will then evaluate the results.

## 6.0 Conclusion

In this paper we have proposed a technique for compiling a bilingual collocational lexicon completely automatically. The techniques use Xtract as a front end in order to identify the collocations to be translated. The translations are then constructed on a word by word basis, and the search space is reduced by only considering words with high mutual information with the original collocation as well as mutually correlated. This paper describes the algorithm and gives some preliminary results. In a next stage, we intend to test the algorithm on more complex cases and then produce a bilingual collocation lexicon to be used by the research community.

BIBLIOGRAPHY

[Bahl *et.al.* 1983] Bahl L., Jelinek F., and Mercer R., "*A maximum likelihood approach to continuous Speech Recognition.*" IEEE transactions on pattern analysis and machine intelligence, 1983, 5-(2), pp:179-190, 1983

[Brown *et.al.* 91a] Brown P. , Della Pietra S., Della Pietra V., and Mercer R. "*Word Sense Disambiguation using Statistical Methods.*" Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics, June 1991, Berkeley Ca.

[Brown *et.al.* 91b] Brown P. , Lai J., and Mercer R. "*Aligning Sentence in Parallel Corpora.*" Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics, June 1991, Berkeley Ca.

[Choueka *et.al.* 83] Choueka Y., Klein T., and Neuwitz E., "*Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus.*" Joùrnal for Literary and Linguistic Computing, vol. 4, pp: 34-38, 1983.

[Gale & Church 91] Gale W., and Church K., "*A program for Aligning Sentences in Bilingual Corpora.*" Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics, June 1991, Berkeley Ca.

[Smadja & McKeown 90] Smadja F., and McKeown K., "*Automatically Extracting and Representing Collocations for Language Generation.*" Proceedings of the 28th Annual Meeting of the Association of Computational Linguistics, June 1990, Pittsburgh, Pa.

[Smadja 92] Smadja F. "*Retrieving Collocations from Text: Xtract.*" Journal of Computational Linguistics, to appear. smadja