Web corpora for bilingual lexicography. A pilot study of English/French collocation extraction and translation

Adriano Ferraresi,* Silvia Bernardini,* Giovanni Picci,** Marco Baroni***

* University of Bologna, Italy

** Larousse dictionnaires bilingues, Paris, France

*** University of Trento, Italy

1. Introduction

This paper describes two very large (> 1 billion words) Web-derived "reference" corpora of English and French, called *ukWaC* and *frWaC*, and reports on a pilot study in which these resources are applied to a bilingual lexicography task focusing on collocation extraction and translation.

The two corpora were assembled through automated procedures, and little is known of their actual contents. The study aimed therefore at providing mainly qualitative evaluation of the corpora by applying them to a practical task, i.e. ascertaining whether resources built automatically from the Web can be profitably applied to lexicographic work, on a par with more costly and carefully-built resources such as the British National Corpus (for English).

The lexicographic task itself was set up simulating part of the revision of an English-French bilingual dictionary. Focusing unidirectionally on English=>French, it first of all compared the coverage of ukWaC vs. the widely used BNC in terms of collocational information of a sample of English SL nodewords. The evidence thus assembled was submitted to a

professional lexicographer who evaluated relevance. The validated collocational complexes selected for inclusion in the revised version were then translated into French drawing evidence from frWaC, and the translations were validated by a professional translator (native speaker of French). The results suggest that the two Web corpora provide relevant and comparable linguistic evidence for lexicographic purposes.

The paper is structured as follows: Section 2 sets the framework for the study, reviewing current approaches to the use of the Web for cross-linguistic tasks, describing the Web corpora used, and the applications of corpora in lexicography work. Section 3 presents the objectives of the pilot investigation, the method followed and its results. In Section 4, we draw conclusions and suggest directions for further work.

2. Corpora, lexicography and the Web

2.1 Web corpora for cross-linguistic tasks

In many fields, ranging from corpus linguistics to Natural Language Processing (NLP) and machine translation (MT), the Web is being increasingly used as a source of linguistic data. This is the case, for instance, when traditional corpus resources prove inadequate to answer certain research questions, either because they are too small and do not contain sufficient evidence for analysis (Kilgarriff and Grefenstette 2003), or because they are not up-to-date enough to document relatively new linguistic phenomena (Brekke 2000). In other cases, e.g. the study of specialized linguistic sub-domains or minority languages, no resource exists (Baroni and Bernardini 2004; Scannel, 2007).

The lack of adequate corpus resources is particularly felt in cross-language studies and NLP, where parallel corpora (originals in language A and their translations into language B) are

often needed but are not available due to the scarcity of relevant (easily and freely accessible) textual materials. In these cases, too, attempts have been made to use the Web as a data source. Resnik and Smith (2003) and Chen and Nie (2000), for example, propose two distinct algorithms to automatically build bilingual corpora from the Web for a variety of language pairs. Their corpora, however, suffer from a number of problems, such as their relatively small size (Resnik and Smith (2003) report that their largest corpus, for the language pair English-Chinese, contains fewer than 3,500 document pairs), and the impossibility to distinguish with certainty which document in a pair is the original and which is the translation.

A more promising approach to using the Web for mining cross-linguistic data is to exploit Web texts to build *comparable* – rather than *parallel* – corpora, and design algorithms that do not require input texts to be one the translation of the other. Drawing on early work by Rapp (1995) and Fung (1995), there is by now a large and growing literature on using "unrelated" (non-parallel) corpora for tasks such as MT and automatic construction of bilingual lexicons (see also Section 4). Witness to this is a workshop organized at the 2008 LREC conference, whose aim was to explore the potential of comparable corpora in tasks for which parallel corpora are traditionally considered the mainstays (Zweigenbaum et al. 2008). The Web was used extensively by the workshop contributors to retrieve (monolingual) corpora for multiple languages sharing similar topics or genres, such as corpora composed of science news texts (Saralegi et al. 2008) or online newspaper texts (Otero 2008).

In the pilot study described in this paper we used two very large, Web-derived corpora of British English (ukWaC) and French (frWaC). Our aim in building them was to set up resources that would be similar, in terms of the variety of text types and topics featured, to more traditional general language corpora (in particular, the British National Corpus, a well-established standard

for British English; to the best of our knowledge, no similar resource exists for French). UkWaC and frWaC thus aim at providing similar "reference" resources for the languages of interest, rather than being comparable to each other by design, as was the case for the corpora used in the experiments discussed above.[1] However, given their large dimensions (>1 billion words), and since they were built following the same procedure, which is described in greater detail in Section 2.2 below, we hypothesize that they could perform comparably in a task whose aim is the extraction of lexicographically relevant information for the languages in question.

2.2 Introducing the WaCky pipeline: frWaC

2.2.1 Introduction

This Section briefly describes the procedure that was followed to construct the corpora used in the experiment. It should be noted that the construction of ukWaC and frWaC follows that of two similar corpora of German (deWaC) and Italian (itWaC) – these resources are among the achievements of an international research project called *WaCky* (***W**eb **a**s **C**orpus **k**ool **y**initiative*).[2] Since the procedure developed within this project (described in detail in Baroni et al. (submitted), and Ferraresi et al. (2008), the latter focusing on ukWaC) is largely language-independent, in this Section attention will be paid especially to those aspects specific to the construction of frWaC. We will focus in particular on the initial steps of the procedure, i.e. "seed" URLs selection and crawling, during which critical language-specific decisions regarding the document sampling strategy are made.

2.2.2 Seed selection and crawling

Our aim was to set up resources comparable to more traditional general language corpora,

containing a wide range of text types and topics. These should include both 'pre-Web' texts of a varied nature that can also be found in electronic format on the Web (spanning from sermons to recipes, from technical manuals to short stories, and ideally including transcripts of spoken language as well), and texts representing Web-based genres (Mehler et al. forthcoming), like personal pages, blogs, or postings in forums. It should be noted that the goal here was for the corpora to be representative of the languages of interest, i.e. (for frWaC) contemporary French in general, rather than representing the French Web.

The first step consisted in identifying sets of seed URLs which would ensure variety in terms of content and genre. In order to find these, around 1,800 random pairs of randomly selected content words were submitted to Google. Previous research on the effects of seed selection upon the resulting Web corpus (Ueyama, 2006) suggested that automatic queries to Google which include words sampled from traditional written sources such as newspapers and reference corpus materials (which typically privilege formal written language) tend to yield 'public sphere' documents, such as academic and journalistic texts addressing socio-political issues and the like. Issuing queries with words sampled from a basic vocabulary list, on the contrary, tends to produce corpora featuring 'personal interest' pages, like blogs or bulletin boards. Since it is desirable that both kinds of documents are included in the corpus, different seed sources were sampled. Two sets of queries were generated: the first set (1,000 word pairs) was obtained by combining mid-frequency content words from a collection of texts published between 1980 and 2000 in the *Le Monde Diplomatique* newspaper. In order to obtain more basic, informal words, the second list of queries (769 word pairs) was generated from a vocabulary list for children from eight to ten years old.[3] The URLs obtained from Google were submitted to the Heritrix crawler[4] in random order, and the crawl was limited to pages in the .fr Web domain

whose URLs do not end in a suffix cuing non-html data (.wav, .jpg, etc.).

2.2.3 Post-crawl cleaning and annotation

The crawled documents underwent various cleaning steps, meant to drastically reduce noise in the data. First, only documents that were of mime type text/html and between 5 and 200 KB in size were kept for further processing. As observed by Fletcher (2004), very small documents tend to contain little genuine text (5KB counting as "very small" because of the html code overhead) and very large documents tend to be lists of various sorts, such as library indices, shop catalogues, etc. We also identified and removed all documents that had perfect duplicates in the collection, since these turned out to be mainly repeated instances of warning messages, copyright statements and the like. While in this way we might also have wasted relevant content, the guiding principle in our Web-as-corpus construction approach is that of privileging precision over recall, given the vastness of the data source.

All the documents that passed this pre-filtering stage underwent further cleaning based on their contents. First, code (html and javascript) was removed, together with the so-called "boilerplate", i.e., following Fletcher (2004), all those parts of Web documents which tend to be the same across many pages (for instance disclaimers, navigation bars, etc.), and which are poor in human-produced connected text. From the point of view of our target user, boilerplate identification is critical, since too much boilerplate will invalidate statistics collected from the corpus and impair attempts to analyse the text by looking at KWiC concordances.

Relatively simple language filters were then applied to the remaining documents, so as to discard documents in foreign languages and machine-generated text, such as that used in pornographic pages to "trick" search engines. Finally, near-duplicate documents, i.e. documents

sharing considerable portions of text, were identified and discarded through a re-implementation of the "shingling" algorithm proposed by Broder et al. (1997).

At this point, the surviving text was enriched with part-of-speech and lemma information, using the TreeTagger.[5] Table 1 gives size data about each stage of the construction of frWaC; the same kind of information is also provided for ukWaC.

	frWaC	**ukWaC**
n of seed word pairs	1,769	2,000
n of seed URLs	6,166	6,528
raw crawl size	470 GB	351 GB
size after document filtering and near-duplicate cleaning	9 GB	12 GB
n of documents after near-duplicate cleaning	2.2 M	2.69 M
size with annotation	27 GB	30 GB
n of tokens	1,027,246,563[6]	1,914,150,197
n of types	3,987,891[6]	3,798,106

Table 1. Size data for frWaC and ukWaC

2.3 Corpus use and dictionary making

The lexicographic industry has always been one of the driving forces behind corpus development, as well as being one of its main beneficiaries. Two of the major corpus building projects of the nineties, leading to well-known and widely used resources like the Bank of English and the British National Corpus, were carried out by academic-industrial consortia in which publishing houses featured prominently, and which saw 'reference book publishing' as the primary application area for the corpora (Burnard 1995). Sinclair's work on the Cobuild dictionaries (described e.g. in Sinclair and Kirby (1990)) shows how corpus informed methods could profitably be applied to obtain information about word and word sense frequency (thus guiding selection from a pre-compiled (e.g. dictionary-derived) headword list), collocation, syntactic patterning, typical usage and typical form (e.g. of a verb). But the corpus also made it to the published Cobuild dictionaries in a more noticeable way, providing not only examples but

also the raw material for the well-known Cobuild definitions (e.g., (**immune**) 'if you are immune to a disease you cannot be affected by it' as opposed to (**immune**) 'Protected from or resistant to some noxious agent or process', OED online). These definitions sometimes also included subtle meaning generalizations unlikely to be obtainable from sources other than the corpus, e.g. typical *semantic prosodies* ((**set in**) 'if something unpleasant sets in it begins and seems likely to continue or develop', underlined added).

Within English lexicography, corpus resources are nowadays generally recognized as indispensable tools of the lexicographer's trade, even by professionals stemming from a non-corpus tradition (see e.g. Landau 2001, ch. 6). Caveats and limitations do remain, of course, both with regard to corpus construction and processing. However large and carefully built, no corpus will ever represent the whole of a language, including its potential for creativity; furthermore, corpora soon become obsolete for the purposes of lexicography, requiring constant updates and enlargements (Landau 2001). In terms of corpus processing, reliance on automation (of corpus annotation and querying) is becoming indispensable as corpora become larger and larger; yet NLP tools (taggers, lemmatizers, parsers) might hide evidence about uncommon or novel usages, while "smart" query tools (Kilgarriff et al. 2008), welcome as they are for speeding up the lexicographer's work, inevitably reduce her control over data selection. Nonetheless there seems to be general consensus that, as claimed by de Schryver (2003: 167), 'no serious compiler would undertake a large dictionary project nowadays without having one (and preferably several) [corpora] at hand'. Interestingly, availability of texts in electronic format through search engines such as Google has not made corpora obsolete, quite the contrary. At the moment, these tools are not sophisticated enough to cope with the needs of linguists (Baroni and Bernardini 2007), and chances are slim that they will ever be, thus making the provision of very large and up-to-date

corpora still a priority for linguistics and the language industry. This is especially true for languages like French, for which large and easily accessible corpus resources are still scarce.

3. Evaluating Web corpora for lexicography: our pilot investigation

3.1 Objectives and method

In the pilot study described in this paper our aim was that of using our automatically-constructed Web corpora for a practical application, namely to derive information about language use for dictionary making/revising. This task can only provide us with indirect evidence about corpus contents and cross-linguistic comparability, yet our take on such issues as quality and representativeness in corpus construction, especially when it gets to large and automatically-constructed corpora, is that the proof of the corpus is in the using. The number of users and usages to which a corpus is put is the ultimate testimony of its scientific as well as practical value, and this applies to automatically- and manually-constructed corpora alike – cf. also the position taken by Atkins et al. (1992: 5) in a seminal paper on the representativeness of (manually constructed) corpora:

> [a]ll samples are biased in some way. Indeed, the sampling problem is precisely that a corpus is inevitably biased in some respects. *The corpus users must continually evaluate the results drawn from their studies and should be encouraged to report them.* (emphasis added)

The task described in this article simulates corpus-based lexicographic practice and combines/contrasts corpus insights with translator/lexicographer input. Collocational information

about three English lexical headwords is collected from ukWaC and submitted to a lexicographer for validation. The validated collocations are then translated with the help of frWaC and with the assistance of a professional translator from English into French. A detailed description of the method follows.

The extraction of potentially interesting English collocational complexes was done in three steps. First, we needed to select words which a lexicographer may want to analyse when dealing with a dictionary revision task. We therefore asked a lexicographer (a native speaker of British English) to provide us with a list of English words (one for each of the three main lexical word classes, i.e. one adjective, one noun and one verb), whose entries might in his opinion be in need of revision in a French/English bilingual dictionary. The words selected by the lexicographer were “hard” (adjective), “point” (noun) and “charge” (verb).

The second step consisted in extracting potentially interesting collocational complexes these headwords may take part in. To do this, we wrote simple rules for the extraction of candidate pairs according to syntactic criteria. While this method has potentially lower recall than one based on simple co-occurrence (i.e., one that disregards syntactic patterning), and is vulnerable to tagging errors, we estimated that precision was to be favoured over recall: since professional lexicographers are typically hard-pressed for time, limiting the amount of noise in the lists was crucial. The idea of overcoming the limitations of “grammatically blind” collocation lists relying on syntactic patterning is also at the basis of the Sketch Engine,[7] a widely used (commercial) corpus query tool especially designed for lexicographic needs. The patterns we chose were:

- for “hard”: all the nouns that occur in a span of one-three words on the right of the adjective;

- for “point”: all the nouns that occur in a span of one-three words on the left of the adjective;
- for “charge”: all the nouns that occur in a span of one-three words on the right of the adjective.

Notice that these grammatical patterns are also used in the Sketch Engine (Kilgarriff et al. 2004) and are among what Atkins et al. (2003: 278) describe as ‘lexicographically relevant sentence constituents’. In all the three cases we extracted lemmas, and did not take into account the words intervening between node and collocate (i.e., we do not distinguish between, e.g., “access point” and “access to this point”) .

The extracted pairs were ranked according to the Log-likelihood measure (Dunning 1993).[8] The top 30 collocational complexes extracted from ukWaC and the BNC were merged into a single list and sorted in alphabetical order. The lexicographer was then asked to look at the three lists and flag the sequences he reckoned might be considered for inclusion in the English part of an English/French bilingual dictionary (whether as a usage example, or a collocation, or anywhere else in the entry), and to provide at his discretion additional comments and observations. Table 2 reports data about the number of word pairs which were sent out to him for evaluation, split by the corpus they were extracted from. The lexicographer analysed the three lists, evaluated their relevance to the specified task, added his comments and returned the files.

Source corpus	**N.**	**%**
ukWaC and BNC (shared pairs)	51	39.6
Only ukWaC	39	30.2
Only BNC	39	30.2
Total	129	100

Table 2. The extracted English collocations.

The second part of the study consisted in finding likely translation equivalents in frWaC for some of the collocational complexes previously validated by the lexicographer. For this task, which was substantially more labour-intensive than the previous one, we focused on the two major senses/uses of the verb "charge" identified by the English lexicographer, roughly corresponding to the following collocate sets:

1. charge -- assault, burglary, connection, conspiracy, crime, fraud, kidnapping, manslaughter, misconduct, murder, offence, possession, rape, sedition, theft, treason
2. charge -- amount, commission, fee, interest, penalty, pound, premium, price, rate, rent, tax, VAT

For the first sense ('bring an accusation against', OED online), two translation equivalents of the node word "charge", namely "inculper de" and "accuser de" were looked up in frWaC, and the 60 most frequent noun collocates in a span of 1-3 words to the right were selected. Out of this list, the most likely potential equivalents of the English noun collocates were selected and submitted for evaluation to an English=> French professional translator (a native speaker of French).

For the second sense, i.e. 'to impose, claim, demand, or state as the price or sum due for anything' (OED online), the method was reversed. The translator was asked to provide equivalents for the collocate nouns ("somme/montant", "commission", "frais", "intérêt", "pénalité", "prime", "price", "taux", "loyer", "taxe/impôt", "TVA").[9] The verb collocates in a span of 1-3 words to the left of these nouns were searched for in frWaC and the 30 most frequent ones were extracted. Potential translation equivalents found in these lists (KWIC concordances

were obtained when in doubt) were then compared with those suggested intuitively by the translator.

3.2 Results

3.2.1 The validated English collocations

The lexicographer analysed the 129 submitted word pairs, put a tick (✓) next to those that he found to have lexicographic relevance, and provided comments about the different ways in which these expressions might be treated in a dictionary. For instance, with reference to the "hard + [noun]" bigrams, he commented that 'Almost every item […] would be an essential inclusion in a (bilingual) dictionary. They are what I consider to be lexicalized, "hard" collocations with independent meanings'. In other cases (e.g. "charge"), he pointed out that most of the submitted pairs would only be included as 'example collocates given under productive sense categories', and provided labels for such potential sense categories, roughly corresponding to Sinclair's (1996) *semantic preferences* (e.g. "charge + [offence: murder, assault, theft,…]"), or commented that a given sequence would probably only be included in larger dictionaries (e.g. "acupuncture point"). In a few cases (about 8% of the total submitted pairs) he was unsure about the lexicographic relevance of the pair, or his intention was unclear (these cases were marked by a **?**). Given that a corpus can play several roles in the making and revising of a dictionary, including signalling semantic prosodies and preferences and providing examples, and that evaluation of relevance is conditional on the specific task at hand, we consider as validated all word pairs for which we had a definite ✓ or a **?**, regardless of accompanying comments (though we do take comments and uncertainties into account in the more qualitative part of the analysis of results).

The results of the expert validation (Table 3) suggest that more than 70% of the word pairs automatically extracted from both the BNC and ukWaC would be potentially relevant for lexicographic purposes, with both corpora contributing very similar numbers of valid collocations. In fact, a slightly higher overall number of valid collocations come from ukWaC than from the BNC (76 vs. 72), even though ukWaC also has a higher number of uncertain cases, which, if factored out, tip the balance in favour of the BNC (69 vs. 67). The similarity in the numeric results obtained from the two corpora is confirmed by the substantial overlap in terms of the actual sequences found. While either corpus contains between 25% and 30% of collocations not found in the other, as many as 45% are present in both lists. These are likely to be the stronger, more time-resistant "core" collocations in the language, e.g.:

- power point, vantage point, melting point
- hard cash, hard hat, hard shoulder
- charge battery, charge offence, charge fee

Source corpus	Yes	Maybe	Selected	%
ukWaC and BNC (shared pairs)	45	1	46	45.0
BNC (not in ukWaC)	24	2	26	25.4
ukWaC (not in BNC)	22	8	30	29.4
ukWaC total	67	9	76	74.5
BNC total	69	3	72	70.5
Total	91	11	102	100
Out of total submitted	129			

Table 3. Results of expert validation

Moving on to an analysis of results broken down by pattern (Table 4), more than 50% of the validated sequences for both the "hard + [noun]" and "[noun] + point" sequences are found in both corpora. Yet the two patterns differ in terms of percentages of valid collocations found only in one or the other corpus. While ukWaC and the BNC have similar numbers of "hard +

[noun]" collocations (26 and 27), this is not the case with "[noun] + point" collocations. As many as 25 out of 28 sequences following this pattern taken from ukWaC are judged to be valid, vs. 18 only from the BNC. This seems mainly due to several occurrences in the BNC list of "[number] point" (e.g. "eleven point", "fourteen point", "nought point", "O point", "twelve point"; notice that this pattern is not attested in the ukWaC list). The "charge + [noun]" pattern is even more interesting in terms of qualitative differences between the two corpora. While similar numbers of valid collocations are found in ukWaC and the BNC, with the BNC performing slightly better than ukWaC (28 vs. 26 collocations), analysis of the actual patterns found in the two corpora and of the lexicographer's comments suggests that the BNC output may in fact be less relevant for lexicographic purposes than that from ukWaC. This is because as many as 15 (out of 30) pairs exemplify the pattern "charge (s.o.) with [offence]". The lexicographer rightly commented that these would only be relevant as examples of the general pattern, but each actual sequence would contribute little to an understanding of word usage, and would certainly not be included as strong collocations. The BNC output thus provides fewer instances of collocations featuring the "charge" ("take as payment") sense of the verb (e.g. "charge + fee, price, VAT, penalty, rent"), and no instance of the pattern "charge + [person]", e.g. "charge customer" found in ukWaC and validated by the lexicographer. More importantly, the only two collocations to get two ticks out of the total submitted (signaling high relevance according to the lexicographer) were found in the ukWaC output for "charge + [noun]", namely: "charge + card" and "charge + zone" (see Concordances 1 and 2). While a few occurrences of "charge + card" do occur in the BNC (7, the numbers are too small to make it to our list), the collocation "(congestion) charging zone" is completely absent from the corpus. As can be seen from Concordance 2., the expression refers to a traffic regulation scheme first implemented in London in 2003, and nowadays operating in

many other cities within and outside the UK. It is not surprising that the expression is absent from the BNC, created in the early nineties, and that the lexicographer found it particularly relevant for purposes of dictionary revision.

Source corpus	Yes	Maybe	Selected	%
Hard				
ukWaC and BNC (shared pairs)	18	1	19	55.8
BNC (not in ukWaC)	7	1	8	23.5
ukWaC (not in BNC)	6	1	7	20.5
ukWaC total	24	2	26	76.4
BNC total	25	2	27	79.4
Total selected	31	3	34	100
Out of (submitted)	41			
Point				
ukWaC and BNC (shared pairs)	15	0	15	53.5
BNC (not in ukWaC)	2	1	3	10.7
ukWaC (not in BNC)	7	3	10	35.7
ukWaC total	22	3	25	89.2
BNC total	17	1	18	64.2
Total selected	24	4	28	100
Out of (submitted)	41			
Charge				
ukWaC and BNC (shared pairs)	13	0	13	31.7
BNC (not in ukWaC)	15	0	15	36.5
ukWaC (not in BNC)	9	4	13	31.7
ukWaC total	22	4	26	63.4
BNC total	28	0	28	68.2
Total selected	37	4	41	100
Out of (submitted)	47			

Table 4. The validated English collocations broken down by pattern

to despatch . We will not <charge your card> until we have confirmed
hed in the UK . We do not <charge credit cards> until goods are avail
. I do n't mind manually <charging credit cards> at all and if I too
and Switch . We will not <charge your card> until your order is disp
booking if less) will be <charged to your card> by the Rowcroft Hote
price you pay . Will you <charge my credit card> when I book ? No ,
N.B. Boys Stuff will not <charge your card> until we are ready to di
to despatch . We will not <charge your card> until we have confirmed
over the phone . We will <charge your credit card> manually . Pre-pa
ment and clothing . Goods <charged by credit card> are normally dispa

Concordance 1. 10 occurrences of "charge + card" from ukWaC

```
ess of the central London <charging zone> has shown that tolls on
. But unlike the existing <charging zone> , there would be no flat
Traffic delays inside the <charging zone> remain 30 % lower than b
ondon or other congestion <charging zones> . Overspeed warning , th
ing the eight square mile <charging zone> . Anyone who enters the
 further extension of the <charging zone> should only be considere
i of synagogue within the <charging zone> , quoted in the Observer
ing within the congestion <charging zone> . This will involve a su
 ride into the congestion <charging zone> . This may be on-street
to pay ? Residents in the <charging zone> can register their vehic
```

Concordance 2. 10 occurrences of "charge + zone" from ukWaC

3.2.2 Translation equivalents from frWaC

With reference to the "bring an accusation against" sense of "charge", a quick browsing of the top 60 noun collocates of "inculper de" and "accuser de" in frWaC shows that 12 out of 16 collocation equivalents of the noun collocates found in ukWaC are present in the output, namely:

- charge burglary ~ inculper vol; accuser vol
- charge connection ~ inculper complicité; accuser complicité
- charge conspiracy ~ inculper conspiration
- charge crime ~ inculper crime; accuser crime
- charge fraud ~ inculper fraude; accuser fraude
- charge manslaughter ~ inculper homicide
- charge murder ~ inculper homicide; inculper meurtre; accuser meurtre
- charge offence ~ inculper délit; accuser délit
- charge possession ~ inculper détention
- charge rape ~ inculper viol
- charge theft ~ inculper vol; accuser vol
- charge treason ~ accuser trahison; inculper trahison

All these translation equivalences were validated by the French translator.

With reference to the second sense, i.e. 'to impose, claim, demand, or state as the price or sum due for anything' (OED online), Table 5. shows the potential equivalents for "charge" in these collocations, found browsing the top 30 verbs co-occurring with each noun in FQ order:

	commission (commission)	droit (duty)	frais (fee)	intérêt (interest)	loyer (rent)	pénalité (penalty)	prime (premium)	prix (price)	somme/ montant (amount)	taux (rate)	taxe/ impôt (tax)	TVA (VAT)
appliquer	*	*				✓	*	*		✓*	✓	✓*
demander	*						✓					
facturer	*		*		*			*	*			✓
faire									✓			
imposer						✓				✓		
infliger						✓						
lever											✓	
payer			✓	✓			✓		✓		✓	✓
percevoir	✓	*			✓						✓	
pratiquer					✓			✓				
prendre	✓											
prélever				*			*				✓*	
recevoir	✓											
réclamer					✓							

Table 5. French equivalents of "charge" ("*demand a sum*")

In Table 5, a tick means that a given verb was found in frWaC among the top collocates for the corresponding noun, while a star means that the verb was given (intuitively) by the French translator as an equivalent of "charge" in that collocation. While a star not accompanied by a tick does not imply that the translator's intuition is faulty,[10] the partial overlap between corpus evidence and translator's intuition does suggest that browsing a large corpus such as frWaC is crucial for several reasons. First, a translator might want to enlarge the pool of possible translation equivalents to be evaluated, rather than relying on her intuition only. Second, certain equivalents of "charge" (e.g. "appliquer" and "percevoir") seem to collocate more widely than others, and would therefore provide safer bets, e.g. for compact dictionaries with limited space. Finally, the presence of "payer" in several lists would seem to suggest that the action of "imposing a price/sum/tax etc. for sthg." might preferably be encoded in French from the perspective of the *imposee*. Both the verbs "faire" and "payer" appear among the collocates as

part of the turn of phrase “faire payer”, and several occurrences of “payer + [noun]” would indeed appear to be translatable into English as “be charged + [noun]”, cf. examples [a]-[c] from frWaC:

[a] Les transportés <paieraient une somme> qui serait un peu supérieure aux tarifs des transports en commun…

[b] Les emprunteurs <paient une commission> de risque , versent souvent une contribution restituable…

[c] … ceux qui ont une activité nécessairement polluante <payent une taxe> et on droit à un allègement ensuite

3.3 Discussion

The results discussed in Section 3.2 suggest that the *WaCky* pipeline, in which texts are collected opportunistically from the Web through automatic routines requiring no manual intervention, produces corpora that compare favourably with high-quality benchmark resources. With reference to the English part of the study, previous attempts at evaluating ukWaC have shown that the corpus is indeed reasonably similar to the BNC in terms of topic coverage (Ferraresi et al. 2008). The current, more functionally-oriented study shows that the two corpora perform comparably in the collocation extraction task, with a slight edge for ukWaC. Besides (obviously) providing more up-to-date results, the latter corpus has a better coverage of different word senses, and does not give undue prominence to uninteresting collocates (e.g. [number] point). While one cannot rule out the possibility that this is an effect of the statistical measure used to extract the collocates, the co-occurrence statistic used, Log-likelihood, is standard in corpus work, and has been shown to provide stable results from corpora of the size of the BNC and smaller (Dunning 1993). The edge in favour of ukWaC is likely to be the result of its substantially larger size, which makes up for the (probably) more reduced variety of texts sampled. Given that constantly updated and very large corpora are required for lexicographic purposes (Landau 2001), and that building a carefully designed corpus like the BNC is very

costly and time-consuming, Web corpora would appear to be a viable alternative.

The results of the second part of our study, using the newly-built French corpus for finding translation equivalents, could not rely on a benchmark for comparison, given that no corpus comparable to the BNC is available in the public domain for the French language. Nonetheless, our study has shown that frWaC provides a very large number of plausible potential collocations that a lexicographer/translator could draw from when translating collocations or examples from English into French. For one of the senses of the verb "charge" (i.e., "bring an accusation against"), an automatic search for two central equivalents finds most of the French noun collocates corresponding to the English noun collocates found in ukWaC. For a second sense of "charge" ("demand a sum"), the opposite method (searching for the verbal collocates of specific nouns) provides a list of verbs that a lexicographer could choose from when translating "charge" collocations, and gives an idea of which verbs have a wider or more restricted range of collocating nouns, and what these are. Comparison between the collocations intuitively suggested by a native speaker translator and those found in the frWaC lists (limited to the top 30 pairs in frequency order) shows that all the verbs the translator came up with are also present in frWaC, though not necessarily in combination with the same nouns. More interestingly perhaps, frWaC suggested that "(faire) payer" is a favourite option that the translator did not come up with, possibly because of its less lexicalized status.

4. Conclusion and further work

This article has introduced two very large corpora of English and French built from the Web following similar procedures, ukWaC and frWaC. Previous studies have shown that evaluating Web corpus contents is an extremely arduous task. This becomes daunting when one attempts to

also compare these contents cross-linguistically; therefore an empirical approach was favoured. A bilingual lexicography task was set up simulating part of a dictionary revision project. English source language collocations were extracted from ukWaC and from a benchmark corpus (the BNC), and validated by a lexicographer. This phase of the task suggested that an automatically built corpus like ukWaC provides results that are both quantitatively and qualitatively similar to those provided by a smaller and older but much more carefully constructed corpus, and (as could be expected) that these results are more useful for a lexicographer because they provide a more up-to-date snapshot of language in use, and because its larger size provides a better coverage of certain word senses. In the bilingual task, the French corpus was used to seek likely translations for the English collocations. This second part of the task was meant to ascertain that the two Web corpora are similarly adequate for practical purposes. Since we are not aware of any publicly available benchmark corpus for French, the paper tried to establish the validity of frWaC by implication: it first showed that ukWaC performs slightly better than the BNC in this task, and then showed that comparable linguistic information can be obtained from ukWaC and frWaC, thus suggesting that frWaC is a valid reference resource for French lexicography, and arguably for the French language in general.

While the results obtained in this study are very encouraging (one should not forget that Web corpora such as those described here are built fully automatically, with no control over corpus contents, and that therefore their validity cannot be assumed *a priori*) a lot remains to be done. We see two main areas in which further work is needed, first to improve on the Web corpora themselves with respect to the requirements of lexicographers, and secondly to investigate the extent to which these new and freely-available resources can be applied to lexicography.

With regard to the first area, in the immediate future we intend to make frWaC available through the Sketch Engine. This software provides users with so-called "word sketches", i.e. summaries of a word's collocational behaviour, generated automatically using rather sophisticated pre-defined syntactic rules. In this paper very simple rules were used for extracting collocations, which only specified the distance between the two co-collocates, and allowed for virtually no flexibility in the searched patterns, thus probably discarding several interesting collocations not matching the search pattern exactly, and including some noise. While leaving the tasks of devising search rules and the need of browsing large amounts of data with the individual user places her in control, we should not forget that '[t]ime pressures too often push the lexicographer to cut corners to avoid time-consuming analyses', and that 'no one will have the time to sort through thousands of hits […] in order to find a particular usage that has to be included' (Landau 2001: 322-323). The trade-off between ease of consultation and control is certainly not news to corpus users, but with Web corpora being constructed opportunistically and reaching sizes of one or more billion words, it is likely to become a major issue in the future, with practical and theoretical implications that need to be explored. In the longer run, we hope to be able to include genre/topic information with the texts in the two corpora. In this sense, one aspect that seems particularly worthy of attention in Web-as-corpus linguistics at large is the development of classificatory schemes that can be applied automatically to Web corpora. Often a lexicographer or translator will need specialized (sub)corpora rather than huge undifferentiated masses of text, e.g. when seeking evidence about specialized senses of a certain word, or when compiling thematic sections. While traditional corpora often contain extra-linguistic information annotated with the texts by the corpus compilers, the creation method used for Web corpora and their size makes manual annotation impossible. Work is therefore underway (see e.g. Sharoff

forthcoming, Biber and Kurjan 2007) to come up with genre/topic typologies adapted to Web genres that can then be used to classify documents within Web corpora using probabilistic classifiers.

Moving on to the second area, i.e. potential applications of multilingual Web corpora to lexicography, we see two main ways in which our corpora can be of immediate relevance. First, in the production of headword lists of currently used single words and phrases, providing suggestions for new inclusions in revised editions of existing dictionaries. Even relatively straightforward methods, e.g. filtering a lemma list from ukWaC using an older corpus as a stoplist, work very well. For instance, the top ten most frequent nouns in ukWaC after filtering out nouns also found in the BNC are: “website”, “dvd”, “websites”, “sudoku”, “linux”, “html”, “Google”, “url”, “blog” and “homepage”. An even simpler procedure, applying no filtering whatsoever and simply listing the most frequent noun-noun sequences in the corpus, provides the following list of high usage potential multiword expressions in English and in French (in the case of French an optional empty slot is allowed between the two nouns):

French phrase	**fq.**	**English phrase**	**fq.**
mot de passe	30379	Web site	175642
chiffre d’affaires	31831	case study	81127
projet de loi	42517	search engine	70514
site Internet	44578	application form	66693
millions d’euros	44901	credit card	65198
prise en charge	48657	Web page	60626
base de donnée	50725	car park	56721
site Web	55954	health care	48833
point de vue	69419	climate change	47655
mise en place	73216	email address	46643

Table 6. Frequent Noun-Noun sequences in ukWaC and frWaC

A lexicographer can quickly browse through these lists to pick expressions that might be included in a revised edition of a dictionary, or whose entries are in need of revision due to their having become key in a given culture (note, e.g., the high frequency of “climate change” in

ukWaC). Secondly, and more challengingly, we intend to investigate the potential of our Web corpora for the automatic extraction of bilingual collocation pairs. In previous work (see Section 2.1), attempts have been made at developing algorithms that find likely translations of single words from relatively small comparable corpora, made of homogeneous classes of textual materials – mainly newspaper texts. Collocational complexes such as those described in this paper, i.e. noun-noun, verb-noun and adjective-noun word pairs, constitute much rarer events in a language than words taken in isolation. For this reason, larger data sets are needed to face the problem of data sparseness in tasks involving automatic extraction of collocations. We would therefore like to test the suitability of corpora like ukWaC and frWaC for such tasks, on the hypothesis that size could compensate for reduced comparability by design. The results presented in Section 3.2 are very encouraging in terms of the comparability of the linguistic evidence obtainable from the two corpora, especially since the likely translations were extracted through computationally unsophisticated methods. Applying a more fine-tuned algorithm on the *WaCky* corpora, we hope to be able to assist lexicographers in the complex task of establishing translation equivalents above the word level.

Acknowledgements

We wish to thank Martyn Back and Christine Gallois for their help with the English collocations and French translations; Federico Gaspari and Sara Piccioni for input on the collocation extraction method, and Sara Castagnoli and Eros Zanchetta for this and for their contribution to the development of frWaC.

Notes

[1] Corpus comparability is far from being a clear-cut notion, especially when corpora contain non-homogeneous classes of textual materials (Kilgarriff 2001), or pertain to different languages (Bernardini and Zanettin 2004). Moreover, ukWaC and frWaC were built with semi-automated procedures (see Section 2.2), thus reducing the possibility to control for the materials that end up in their final set up. Given these difficulties, in this paper we do not attempt to provide an evaluation of the similarity between the two corpora in terms of their contents, but rather try to establish whether they can provide comparable resources in the framework of a practical task, chosen among those most central to corpus linguistics (see Section 2.3).

[2] http://wacky.sslmit.unibo.it/

[3] http://o.bacquet.free.fr/index.html

[4] http://crawler.archive.org/

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] These data refer to the beta version of frWaC which is available at the time of writing. We expect that after further processing token and type counts will stabilize to numbers similar to those pertaining to ukWaC.

[7] http://www.sketchengine.co.uk

[8] In order to decide on the best measure to use, a distinct mini-pilot study was conducted. Three lists displaying the top 100 collocational complexes of the noun “course” in ukWaC were first sorted according to bare Frequency (FQ), Log-Likelihood (LL) and Mutual Information (MI). Seven expert linguist informants were then asked to judge what list best fit their intuitions about the collocations of "course" (the word was picked opening a monolingual dictionary at a random page). Four people favoured the LL list, two the FQ list and only one the MI list. Based on these results, and on previous work on collocation extraction (cf. Evert 2008), the LL measure was adopted.

[9] In this search we ignore one of the collocates validated by the English lexicographer, namely “pound”, since this has no obvious equivalent in French (“livre sterling”?, “euro”?, “franc”?).

[10] Note that bare frequency was considered here, and only the top most frequently co-occurring verbs were analysed; it is quite likely that other relevant verbs would turn up if one browsed a longer collocate list, and/or if more

sophisticated co-occurrence statistics were used.

References

Atkins, S., Clear, J. and Ostler, N. (1992) 'Corpus design criteria'. *Literary and Linguistic Computing* 7(2): 1-16.

Atkins, S., Fillmore, C.J. and Johnson, C.R. (2003) 'Lexicographic relevance: selecting information from corpus evidence'. *International Journal of Lexicography* 16(3): 251-80.

Baroni, M. and Bernardini, S. (2004) 'BootCaT: Bootstrapping corpora and terms from the Web', in *Proceedings of LREC*. Lisbon, 1313-16.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (submitted) 'The *WaCky* Wide Web: a collection of very large linguistically processed web-crawled corpora", submitted to *Language Resources and Evaluation.*

Bernardini, S. and Zanettin, F. (2004) 'When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals', in A. Mauranen and P. Kuyamaki (eds.) *Translation Universals: Do they Exist?*, 51-62. Amsterdam: Benjamins.

Biber, D. and Kurjan, J. (2007) 'Towards a taxonomy of Web registers and text types: A multidimensional analysis', in M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*, 109-31. Amsterdam: Rodopi.

Brekke, M. (2000) 'From the BNC towards the Cybercorpus: a quantum leap into chaos?', in J.M. Kirk (ed.) *Corpora Galore: Analyses and Techniques in Describing English*, 227-47. Amsterdam: Rodopi.

Broder, A., Glassman, S., Manasse, M. and Zweig, G. (1997) 'Syntactic clustering of the Web', in *Proceedings of the Sixth International World Wide Web Conference*. Santa Clara (CA), 391–404.

Burnard, L. (1995) *Users Reference Guide for the British National Corpus*. Oxford: OUCS.

de Schryver, G.-M. (2003) 'Lexicographers' dreams in the electronic-dictionary age'. *International Journal of Lexicography* 16(2): 143-99.

Chen, J. and Nie, J.-Y. (2000) 'Parallel Web text mining for cross-language information retrieval', in *Recherche d'Informations Assistée par Ordinateur (RIAO)*. Paris, 62–77.

Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* 19(1): 61-74.

Evert, S. (2008) 'A lexicographic evaluation of German adjective-noun collocations', in *Proceedings of the Workshop on 'Towards a Shared Task for Multiword Expressions' at LREC*. Marrakech, 3-6.

Ferraresi, A., Zanchetta, E., Baroni, M. and Bernardini, S. (2008) 'Introducing and evaluating ukWaC, a very large Web-derived corpus of English', in *Proceedings of the WAC4 Workshop at LREC*. Marrakech, 45-54.

Fletcher, W. (2004) 'Making the Web more useful as a source for linguistic corpora'. In U. Connor and T. Upton (eds.) *Corpus Linguistics in North America 2002*, 191-205. Amsterdam: Rodopi.

Fung, P. (1995) 'Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus', in *Proceedings of the 3rd Annual Workshop on Very Large Corpora*. Boston (MA), 173-83.

Kilgarriff, A. (2001) 'Comparing corpora'. *International Journal of Corpus Linguistics* 6(1): 97–133.

Kilgarriff, A. and Grefenstette, G. (2003) 'Introduction to the special issue on the Web as corpus'. *Computational Linguistics* 29(3): 1-15.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. and Rychlý, P. (2008) 'GDEX:

automatically finding good dictionary examples in a corpus', in *Proceedings of Euralex.* Barcelona, 425-32.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004) 'The Sketch Engine', in *Proceedings of Euralex*. Lorient, 105-16.

Landau, S. (2001) *Dictionaries: The Art and Craft of Lexicography*. Cambridge: CUP.

Manning, C., and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge (MA): MIT Press.

Mehler, A., Sharoff, S., Rehm, G. and Santini, M. (eds.) (forthcoming) *Genres on the Web: Computational Models and Empirical Studies.* University of Bielefeld.

Otero, P.G. (2008) 'Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora', in P. Zweigenbaum, E. Gaussier and P. Fung (eds.) *Proceedings of the Workshop on Comparable Corpora at LREC*. Marrakech, 19-26.

Rapp, R. (1995) 'Identifying word translations in non-parallel texts', in *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*. Cambridge (MA), 320-22.

Resnik, P. and Smith, N. (2003) 'The Web as a parallel corpus'. *Computational Linguistics* 29 (3): 349-380.

Saralegi, X., San Vicente, I., and Gurrutxaga, A. (2008) 'Automatic extraction of bilingual terms from comparable corpora in a popular science domain', in P. Zweigenbaum, E. Gaussier and P. Fung (eds.) *Proceedings of the Workshop on Comparable Corpora at LREC*. Marrakech, 27-32.

Scannel, K.P. (2007) 'The Crúbadán Project: corpus building for under-resourced languages', in C. Fairon, H. Naets, A. Kilgarriff and G.-M. de Schryver (eds.) *Building and Exploring Web corpora. Proceedings of the WAC3 Conference*. Louvain, 5-15.

Sharoff, S. (forthcoming) 'In the garden and in the jungle: comparing genres in the BNC and

Internet', in A. Mehler, S. Sharoff, G. Rehm and M. Santini (eds.) *Genres on the Web: Computational Models and Empirical Studies.* University of Bielefeld.

Sinclair, J. McH. and Kirby, D. (1990) 'Progress in English computational lexicography'. *World Englishes* 9(1): 21-36.

Sinclair, J. McH. (1996) 'The search for units of meaning'. *Textus*, 9(1): 71-106.

Zweigenbaum, P., Gaussier, E., and Fung, P. (eds.) (2008) *Proceedings of the Workshop on Comparable Corpora at LREC 2008*. Marrakech, June.