[*Translating and the Computer 21*. Proceedings… 10-11 November 1999 (London: Aslib)]

# The World Wide Web as a Resource for Example-Based Machine Translation Tasks

Gregory Grefenstette

*Xerox Research Centre Europe, Grenoble, France*

Abstract The WWW is two orders of magnitude larger than the largest corpora. Although noisy, web text presents language as it is used, and statistics derived from the Web can have practical uses in many NLP applications. For this reason, the WWW should be seen and studied as any other computationally available linguistic resource. In this article, we illustrate this by showing that an Example-Based approach to lexical choice for machine translation can use the Web as an adequate and free resource.

Key Words: WWW, Example-Based, machine translation, corpus linguistics, very large lexicon

## 1. Introduction

The idea of using attested linguistic events to choose between theoretically possible events underlies Example-Based Natural Language Processing tasks. This approach has been used for Machine Translation (Sato and Nagao, 1990; Dagan et al, 1991; Sumita et al, 1993) and to improve Cross-Language Information Retrieval (Ballesteros and Croft, 1998). For these tasks, candidate multiword translations are generated using human-compiled electronic dictionaries or using equivalence lexicons derived from bilingual aligned corpora (Brown et al, 1990). The candidate translations are scored using statistics of the candidates' attested appearances in a reference corpus, and the highest scoring candidate are chosen as the translation term.

It is evident that the World Wide Web can be considered as an extremely large corpus of attested examples. Some linguists cringe at the idea of using this uncharacterized and dirty corpus to derive linguistic information, but we argue that the sheer size of the WWW as a corpus allows signal to overcome noise. There exist a few large corpora that have been collected and cleanly prepared, such as the British National Corpus[1] of 100 million words (90 million from written text, and 10 million from spoken text), but the quantity of text available through the Web swamps these collections. To get an idea of the size of the World Wide Web, we show, in Table 1, a list of counts of some random noun

[1] http://info.ox.ac.uk/bnc

phrases in this large British National Corpus and their counts in an indexed Web browser, *AltaVista*[2], on a given day in late 1998.

These examples show that the number of attestable patterns is almost two orders of magnitude larger on the Web than the number to be found in the largest available corpora. Statistical techniques, such as Example-Based methods, rely on the presence of events of to perform well. Many Example-Based techniques suffer performance drop-offs when they try to make choices using rare events, since the distinction between signal and noise becomes blurred. The size of the Web, however, weakens[3] the effect of Zipf's law (Zipf, 1965), since intuitively likely events do become common enough for statistical techniques to work.

| | *BNC* | *WWW* |
|---|---|---|
| *sample phrases* | 100 M Words | |
| medical treatment | 202 | 46064 |
| prostate cancer | 28 | 40772 |
| deep breath | 374 | 54550 |
| acrylic paint | 20 | 7208 |
| perfect balance | 28 | 9735 |
| presidential election | 74 | 23745 |
| electromagnetic radiation | 24 | 17297 |
| powerful force | 54 | 17391 |
| concrete pipe | 8 | 3360 |
| upholstery fabric | 5 | 3157 |
| vital organ | 30 | 7371 |

**Table 1. Counts of some random noun phrases in the British National Corpus and as found on the World Wide Web by the AltaVista browser in late 1998.**

As an anecdotal example of how the Web can be used as a resource in the Example-Based task of lexical choice in dictionary-based machine translation, consider the following example. Take the compositional French noun phrase *groupe de travail.* In the Oxford-Hachette French-English dictionary, the French word *groupe* can be translated by the English words *cluster, group, grouping, concern* and *collective.* The French word *travail* can be translated by the English words *work, labor*

[2] http://www.AltaVista.com

[3] This not to say that noise does not exist, or that every linguistic utterance appearing on the Web is immediately validated by its simple presence. For example, the canonical counter-example of "colorless green"' can be found 337 times via AltaVista. But now that *valid utterances* do occur thousands of times on the Web, the impact of such self-reference generated noise is diminished.

or *labour*. The naïve translator has five (from *groups)* times three (from *travail)* possible ways of translating *groupe de travail*. Now, the AltaVista search portal allows the Web browser user to search for adjacent phrases by placing their query in double-quotes. Combining the possible translations of *groupe de travail* into all twenty-one possible noun phrases creatable by simply re-ordering the nouns and concatenating them to form English phrases, and then submitting these phrases to this Web browser yields, in Table 2, the actual occurrence statistics in the web pages indexed by this browser. We see that the phrase *work group* is much more frequent than all the others, and is the most likely domain-independent translation in the group[4].

| | *WWW count* | | *WWW count* |
|---|---|---|---|
| labor grouping | 4 | labor cluster | 7 |
| labour concern | 8 | work grouping | 27 |
| labor concern | 28 | work cluster | 112 |
| labor collective | 144 | labour collective | 158 |
| work concern | 170 | work collective | 242 |
| labor group | 844 | labour group | 1131 |
| work group | 67238 | | |

**Table 2. Web counts of some possible ways of translation the French expression *groupe de travail* using the possible translations of *groupe* and *travail* given in a bilingual French-English dictionary Some possibilities (eg labour cluster) did not appear at all.**

Going from anecdote to experimentation, we test in the next section the use of the World Wide Web as a resource for Example-Based Machine Translation on a large-scale.

## 2. Experimentation

In order to perform an objective, large-scale experiment on the adequacy of the World Wide Web as a linguistic resource for an Example-Based Machine Translation task, we created a gold standard of compositional compounds from a publicly available electronic bilingual dictionary[5].

---

[4] Though the morphological variant *working group,* found 530124 times is the preferred (as well as the more frequently occurring) translation.

[5] We used the Basic Multilingual Lexicon http://www.icp.qrenet.fr/ELRA/cata/text_det.html#basmullex, available from the ELRA as our dictionary. This dictionary contains 37,600 senses translated across five languages: English, French, Spanish, Italian, and German. We used the German-English and Spanish-English parts.

The standard was created by eliminating all phrases in the dictionary which were not transparent translations of their subparts. We tested two language directions: German-to-English and Spanish-to-English. To find compositional noun phrases in this multilingual dictionary, we extracted two complete sets of all German compound nouns and all Spanish nominal phrases satisfying the four criteria:

i. [compound] the dictionary entry was decomposable into two other Spanish or German words found in the dictionary,
ii. [compositionality] the compound term was translated in the English part of the dictionary by two word phrases,
iii. [transparency] the words in the English translations of the smaller German or Spanish components permitted the construction of candidate translations that included the dictionary-given compound-word translation, and
iv. [ambiguity] there was more than one possible English translation candidate.

These sets of words, then, correspond to the entire list of German compounds and Spanish terms in this full-size dictionary such that, if they were not in the dictionary, their proper English translation could be constructed from the translation of the subparts of the German word or Spanish term using that same dictionary. Only such words which had ambiguous translations were retained. This strategy led to a set of 724 German words constituting our gold standard of potentially ambiguous compositional German compounds, and a set of 1140 compositional Spanish terms. With each German word or Spanish term, we also have their preferred[6] English translations.

For each German word and for each Spanish term, we then ignored the dictionary entry for the compound, and created the English candidate translations as if the non-English term were not included in the dictionary. This situation reproduces what human users must do for most novel German compounds or novel Spanish terms encountered. In each case, we created all the possible two word translations using the decomposed[7] German word and the individual words of the Spanish terms (ignoring prepositions) and recombining the English translations of these subparts from the German-to-English or Spanish-to-English sides of the same dictionary.

---

[6] By preferred, we mean what our dictionary gives as a translation of the term. One might raise the question about whether the dictionary might be wrong in this sense, but to remain objective, we considered that the dictionary was always right.

[7] Decomposed using techniques described in (Schiller, 1996).

Since each of the 724 German compound words was ambiguously translatable (given the translations of their components in the reference dictionary), 3556 possible English translations were generated. For the 1140 ambiguous Spanish multiword terms, there were 6186 possible English translations built using this simple concatenation strategy. Each possible translation candidate was sent to AltaVista as a phrasal query, and the frequency[8] of occurrence of the phrase was noted. To use the WWW as a decision mechanism for choosing the proper translation, the most frequently occurring phrase was chosen as the best example for translating the ambiguous term. This choice was compared against the actual translation that the dictionary gave for them. The results of this experiment are shown in the Table 3, showing that 86-87% of the choices were correct.

| | |
|---|---|
| Number of German nouns responding to 4 criteria | 724 |
| Number of candidate English translations | 3556 |
| Number of correct translations choosing most frequent phrase in AltaVista as best | 631 |
| Percent of correctly chosen translations | 87% |

| | |
|---|---|
| Number of Spanish terms responding to 4 criteria | 1140 |
| Number of candidate English translations | 6186 |
| Number of correct translations choosing most frequent phrase in AltaVista as best | 976 |
| Percent of correctly chosen translations | 86% |

**Table 3. The results of creating translation candidates from subparts of German compounds and Spanish multiword expressions, and then choosing the translation candidate that appears most often in a Web Browser.**

Here are some example of the translation candidates and their AltaVista frequencies. In the following tables, we give some examples of the German compound words and the Spanish terms with the English candidate translations that were generated by translating the components. For each candidate, the number of times that AltaVista had found the phrase is given as *AltaVista count.* The next two columns show whether the frequency information is sufficient to pick a dictionary-given translation: if there is the abbreviation DICT in column 5 then the

[8] The page frequency. AltaVista returns a count of the number of times that a word or expression (enclosed in quotes), has been seen on the pages that it indexes, and the number of WWW pages that contain the term. The counts given in this paper were calculated in the beginning of 1999, and correspond to the number of pages found.

English candidate translation of the components corresponds to the gold standard dictionary translation of the German compound or the Spanish term. The word MAX in the last column shows which of the English candidates was most frequent on the Web indexed by Altavista[9]. 87% of the ambiguous German words and 86% of the ambiguous Spanish multiword terms tested had DICT in column 5 and the word MAX in column 6, meaning that the most frequent attested candidate on the Web was also a gold standard translation of the compound word. For example *Appartementhaus* generates 8 candidate translations: *apartment chop, apartment cut, apartment house, ...* of which *apartment house* is the most common on the Web and the translation given for the compound. On the other hand, *Aktienkurs* generated 8 translations of which *stock price* was the most common but not given in the dictionary. This last example was counted among the 13% incorrect German cases. Notice in the tables that many candidates that are not the most frequent ones still have no-zero frequencies, for example *apple sap*, one of the candidate translations of *Apfelsaft* still appeared 25 times on the Web.

| ***German compound*** | ***English candidate*** | ***AltaVista count*** | ***gold standard*** | ***most frequent*** |
|---|---|---|---|---|
| Angebotspreis | offer price | 9767 | DICT | MAX |
| Angebotspreis | offer prize | 206 | - | |
| Apfelkraut | apple herb | 167 | - | MAX |
| Apfelkraut | apple syrup | 159 | DICT | |
| Apfelsaft | apple juice | 13841 | DICT | MAX |
| Apfelsaft | apple sap | 25 | - | |
| Appartementhaus | apartment chop | 0 | - | |
| Appartementhaus | apartment cut | 127 | - | |
| Appartementhaus | apartment house | 8356 | DICT | MAX |
| Appartementhaus | apartment rampage | 0 | - | |
| Appartementhaus | flat chop | 10 | - | |
| Appartementhaus | flat cut | 621 | - | |
| Appartementhaus | flat house | 882 | - | |
| Appartementhaus | flat rampage | 0 | - | |
| Bogenbrücke | arch bridge | 2304 | DICT | MAX |
| Bogenbrücke | bow bridge | 224 | - | |

An example from the Spanish data shows that this experiment only gives the most common translations (corresponding to those appearing in the

[9] Recent tests from June 1999 estimate that AltaVista indexes about 15% of the static Web pages accessible on the Web.

bilingual gold standard dictionary) whereas in a specific domain, a rarer translation might be acceptable. For example, the experiment erroneously chooses *energy field* as the translation of *campo de fuerzas*, rather than the dictionary supplied *force field*, but the choice of one or the other may well depend on the domain or context of application. Here, we are simply saying that the WWW provides an idea of the most common way of saying something.

| *German compound* | *English candidate* | *AltaVista count* | *gold standard* | *most frequent* |
|---|---|---|---|---|
| Aktienkurs | share course | 246 | - | |
| Aktienkurs | share cure | 0 | - | |
| Aktienkurs | share price | 48221 | DICT | |
| Aktienkurs | share rate | 598 | - | |
| Aktienkurs | stock course | 60 | - | |
| Aktienkurs | stock cure | 5 | - | |
| Aktienkurs | stock price | 48394 | - | MAX |
| Aktienkurs | stock rate | 167 | - | |
| Blutspender | bleed donor | 0 | - | |
| Blutspender | bleed giver | 0 | - | |
| Blutspender | blood donor | 5432 | DICT | MAX |
| Blutspender | blood giver | 5 | - | |
| Blutzelle | bleed cell | 0 | - | |
| Blutzelle | blood cell | 25514 | DICT | MAX |
| Braunkohle | brown cabbage | 20 | - | |
| Braunkohle | brown coal | 2317 | DICT | MAX |
| Briefwaage | letter balance | 509 | DICT | MAX |
| Briefwaage | letter Libra | 2 | - | |
| Briefwaage | letter scales | 131 | DICT | |
| Brotmesser | bread knife | 1167 | DICT | MAX |
| Brotmesser | bread meter | 0 | - | |
| Brotmesser | loaf knife | 0 | - | |
| Brotmesser | loaf meter | 0 | - | |

Note that AltaVista does not index noun phrases but merely contiguous words. These AltaVista counts are a rough estimate of a given noun phrase. This experiment could also be made more subtle by generating more varied syntactic forms (such as *A of B)* or through a more intelligent use of morphological variants, without modifying the way that the available Web browser indexes its pages. Ideally, the Web browsers would perform a more intelligent indexing, extracting not only

contiguous terms but dependency structures that can be derived through current robust, shallow parsing systems (Appelt et al, 1993; Ait-Moktar and Chanod, 1997; Grefenstette, 1997). But even in its simple state, this German and Spanish to English experiment shows that the WWW is a linguistic resource of the same nature and same (though possibly greater) utility as those corpora now used in Natural Language Processing tasks.

## 3. Conclusion and Perspectives

We have presented an experiment in Example-Based Natural Language Processing using the World Wide Web as the exemplar linguistic resource for decision making. Our experiment was on a much larger scale than previous efforts (Dagan et al, 1991; Rackow et al, 1992), limited to a few dozen words, since we included all the potentially ambiguous compounds in a large translation dictionary, and worked with a corpus (the entire WWW visited by AltaVista) that is orders of magnitude larger than any previously used corpus.

| *Spanish term* | *English candidate* | *AltaVista count* | *gold standard* | *most frequent* |
|---|---|---|---|---|
| agregado de prensa | press attaché | 403 | DICT | MAX |
| agregado de prensa | squeezer attaché | 0 | - | |
| agua corriente | common water | 2815 | - | |
| agua corriente | current water | 5213 | - | |
| agua corriente | draft water | 1438 | - | |
| agua corriente | draught water | 11 | - | |
| agua corriente | flowing water | 13264 | - | |
| agua corriente | going water | 343 | - | |
| agua corriente | ordinary water | 2040 | - | |
| agua corriente | power water | 12695 | - | |
| agua corriente | running water | 49358 | DICT | MAX |
| agua corriente | stream water | 9264 | - | |
| agua corriente | usual water | 1252 | - | |
| agua mineral | mineral water | 33058 | DICT | MAX |
| agua mineral | ore water | 178 | - | |
| agua salada | pickle water | 284 | - | |
| agua salada | salt water | 98690 | DICT | MAX |
| águila real | actual eagle | 60 | - | |
| águila real | essential eagle | 11 | - | |
| águila real | real eagle | 176 | - | |
| águila real | royal eagle | 431 | DICT | MAX |
| ahorro de energía | decisiveness saving | 0 | - | |
| ahorro de energía | energy saving | 140148 | DICT | MAX |

A human (or computer) deciding on the correct translation of compositional noun phrases would be faced with the same choice as that presented in this Example-Based Natural Language Processing experiment. An extremely simple exploitation of the WWW provides the linguistic resource, a relatively free resource one might add, to resolve this choice with 86-87% accuracy.

This experiment argues for a greater exploitation and study of the Web as a linguistic resource, and for applying techniques of shallow parsing to create more linguistically informed indexes than those available through current web portals.

## References


Salah Ait-Mokhtar and Jean-Pierre Chanod. 1997. Incremental finite-state parsing. In *ANLP'97,* pages 72-79, Washington.

Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, and Mabry Tyson. 1993. FASTUS: A finite-state processor for information extraction from real-word text. In *Proceedings IJCAI '93,* Chambery, France, August.

Lisa Ballesteros and W. Bruce Croft. 1998. Resolving ambiguity for cross-language retrieval. In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 64-71, Melbourne, Australia, August. ACM Press, New York.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, and Paul S. Roossin. 1990.A statistical approach to language translation. *Computational Linguistics,* 16(2):79—85.

I. Dagan, I. Atai, and U. Schwall. 1991. Two languages are better than one. In *Proceedings of the 29th Meeting of the ACL,* pages 130—137, Berkeley.

Gregory Grefenstette. 1997. SQLET : Short query linguistic expansion techniques: Palliating one or two-word queries by providing intermediate structure to text. In *RIAO'97, Computer-Assisted Information Searching on the Internet,* Montreal, Canada.

U. Rackow, I. Dagan, and U. Schwall. 1992. Automatic translation of noun compounds. In *Proceedings of COLING'92,* pages 1249-1253, Nantes, France, August 23-28.

S. Sato and M. Nagao. 1990. Towards memory-based translation. In H. Karlgren, editor, Proceedings of *COLING'90,* pages 247-252, Helsinki.

Anne Schiller. 1996. Deutsche flexions- und kompositionsmorphologie mit pc-kimmo. In Roland Hausser, editor, *Linguistische Verifikation: Documentation zur Ersten Morpholympics 1994,* number 34 in *Sonderdruck aus Sprache und Information.* Max Niemeyer Verlag, Tubingen.

E. Sumita, K. Oi, O. Furuse, H. Iida, T. Higuchi, N. Takahashi, and H. Kitano. 1993. Example-Based machine translation on massively parallel processors. In *Proc. of the I3th IJCAI,* pages 1283-1288, Chambery, France.

G. K. Zipf. 1965. *Human Behavior and the Principle of Least Effort.* Hafner, New York.

| *Spanish term* | *English candidate* | *AltaVista count* | *gold standard* | *most frequent* |
|---|---|---|---|---|
| ala delta | delta nostril | 0 | - | |
| ala delta | delta wing | 1525 | DICT | MAX |
| ala delta | delta winger | 1 | - | |
| àlbum de sellos | seal album | 56 | - | |
| àlbum de sellos | stamp album | 1805 | DICT | MAX |
| alfombra oriental | easterly carpet | 0 | - | |
| alfombra oriental | eastern carpet | 115 | - | |
| alfombra oriental | oriental carpet | 5985 | DICT | MAX |
| alumbrado de emergencia | emergency lighting | 17940 | DICT | MAX |
| alumbrado de emergencia | emergency lit | 5 | - | |
| ambiente del trabajo | labor atmosphere | 105 | - | |
| ambiente del trabajo | labor cosiness | 0 | - | |
| ambiente del trabajo | labor coziness | 0 | - | |
| ambiente del trabajo | labor snugness | 0 | - | |
| ambiente del trabajo | labour atmosphere | 4 | - | |
| ambiente del trabajo | labour cosiness | 0 | - | |
| ambiente del trabajo | labour coziness | 0 | - | |
| ambiente del trabajo | labour snugness | 0 | - | |
| ambiente del trabajo | work atmosphere | 3437 | DICT | MAX |
| ambiente del trabajo | work cosiness | 0 | - | |
| ambiente del trabajo | work coziness | 0 | - | |
| ambiente del trabajo | work snugness | 0 | - | |
| campaña de propaganda | propaganda campaign | 4337 | DICT | MAX |
| campaña de propaganda | propaganda expedition | 2 | - | |
| campaña publicitaria | advertising campaign | 70816 | DICT | MAX |
| campaña publicitaria | advertising expedition | 3 | - | |
| campaña publicitaria | advertizing campaign | 150 | DICT | |
| campaña publicitaria | advertizing expedition | 0 | - | |
| campeón mundial | world champion | 143343 | DICT | MAX |
| campeón mundial | worldwide champion | 868 | - | |
| campeonato mundial | world championship | 121676 | DICT | MAX |
| campeonato mundial | worldwide championship | 53 | - | |
| campo de concentración | concentration camp | 26532 | DICT | MAX |
| campo de concentración | concentration country | 19 | - | |
| campo de concentración | concentration countryside | 0 | - | |
| campo de concentración | concentration field | 575 | - | |
| campo de concentración | concentration provinces | 0 | - | |

| *Spanish term* | *English candidate* | *AltaVista count* | *gold standard* | *most frequent* |
|---|---|---|---|---|
| campo de fuerzas | energy camp | 769 | - | |
| campo de fuerzas | energy country | 451 | - | |
| campo de fuerzas | energy countryside | 6 | - | |
| campo de fuerzas | energy field | 20968 | - | MAX |
| campo de fuerzas | energy provinces | 8 | - | |
| campo de fuerzas | force camp | 920 | - | |
| campo de fuerzas | force country | 292 | - | |
| campo de fuerzas | force countryside | 3 | - | |
| campo de fuerzas | force field | 16390 | DICT | |
| campo de fuerzas | force provinces | 21 | - | |
| campo de fuerzas | power camp | 103 | - | |
| campo de fuerzas | power country | 501 | - | |
| campo de fuerzas | power countryside | 10 | - | |
| campo de fuerzas | power field | 3301 | - | |
| campo de fuerzas | power provinces | 83 | - | |
| campo de fuerzas | strength camp | 515 | - | |
| campo de fuerzas | strength country | 259 | - | |
| campo de fuerzas | strength countryside | 0 | - | |
| campo de fuerzas | strength field | 556 | - | |
| campo de fuerzas | strength provinces | 7 | - | |
| campo de fuerzas | vigor camp | 1279 | - | |
| campo de fuerzas | vigor country | 29 | - | |
| campo de fuerzas | vigor countryside | 2 | - | |
| campo de fuerzas | vigor field | 97 | - | |
| campo de fuerzas | vigor provinces | 0 | - | |
| campo de fuerzas | vigour camp | 73 | - | |
| campo de fuerzas | vigour country | 1 | - | |
| campo de fuerzas | vigour countryside | 0 | - | |
| campo de fuerzas | vigour field | 3 | - | |
| campo de fuerzas | vigour provinces | 0 | - | |
| campo de fuerzas | violence camp | 1259 | - | |
| campo de fuerzas | violence country | 369 | - | |
| campo de fuerzas | violence countryside | 0 | - | |
| campo de fuerzas | violence field | 179 | - | |
| campo de fuerzas | violence provinces | 4 | - | |

| *Spanish term* | *English candidate* | *AltaV count* | *Gold stand* | most freq |
|---|---|---|---|---|
| campo de fútbol | football camp | 4899 | - | |
| campo de fútbol | football country | 199 | - | |
| campo de fútbol | football countryside | 4 | - | |
| campo de fútbol | football field | 27967 | DICT | MAX |
| campo de fútbol | football provinces | 0 | - | |
| campo de fútbol | soccer camp | 4437 | - | |
| campo de fútbol | soccer country | 114 | - | |
| campo de fútbol | soccer countryside | 1 | - | |
| campo de fútbol | soccer field | 13944 | - | |
| coleccionista de monedas | coin collector | 7165 | DICT | MAX |
| coleccionista de monedas | currency collector | 255 | - | |
| coleccionista de sellos | seal collector | 24 | - | |
| coleccionista de sellos | stamp collector | 8655 | DICT | MAX |
| collar de perlas | pearl collar | 94 | - | |
| collar de perlas | pearl necklace | 9234 | DICT | MAX |
| color de camuflaje | camouflage color | 236 | - | |
| color de camuflaje | camouflage colour | 272 | DICT | |
| color de camuflaje | camouflage paint | 617 | - | MAX |
| columna conmemorativa | commemorative column | 37 | - | |
| columna conmemorativa | commemorative pillar | 18 | - | |
| columna conmemorativa | memorial column | 128 | DICT | MAX |
| columna conmemorativa | memorial pillar | 74 | - | |