A corpus study on the number of true proportional analogies between chunks in two typologically different languages

Yves Lepage, Julien Migeot, Erwan Guillerm

A corpus study on the number of true proportional analogies between chunks in two typologically different languages

Yves Lepage[1] Julien Migeot[2] Erwan Guillerm[2]

[1,2]GREYC, University of Caen, France
Tel. + 33-2-31-56-74-82, Fax. +33-2-31-25-73-30
[1]Yves.Lepage@info.unicaen.fr [2]{jmigeot, eguiller}@etu.info.unicaen.fr

Abstract

We measure the number of true proportional analogies between chunks in two typologically different languages on a similar corpus: a 20,000 sentence long Japanese-English bicorpus. We verify that at least 96% of analogies of form between chunks are also analogies of meaning. We conclude that analogy ought to be considered as a reliable structuring device between chunks.

1 Introduction

The purpose of this paper is to establish that true analogies between chunks in two typologically different languages, English and Japanese, are important in number. More precisely, we show that analogies of form correspond to analogies of meaning with a very high confidence. Such a result gives theoretical support to other works in natural language processing that already make use of analogy.

The paper is structured as follows. Section 2 will refer to previous works in natural language processing and to the current situation in linguistics. Section 3 will introduce the purpose of the paper and recall what 'true' analogies are. Section 4 will describe our experimental protocol and data, and Section 5 the actual experiments. Results will be summarized and commented in Section 6.

2 The situation in natural language processing and linguistics

Proportional analogy is a relationship between four pieces of language, A, B, C and D, that states that 'A is to B as C is to D'. The notation is $A : B :: C : D$. For instance between words (1), sentences (2) or chunks (3):

(1) *unrelated* : *relate* :: *unmodulated* : *modulate*

(2) *Do you like music?* : *Do you go to concerts often?* :: *Do you like classical music* : *Do you go to classical concerts often?*

(3) *my room key* : *the room key* :: *my first visit* : *the first visit*

Recently, a number of works in natural language processing make use of proportional analogies. (Turney and Littman, 2005) show the use of different machine techniques to answer SAT tests (analogical puzzles on words) with scores comparable to those of a human being; (Stroppa and Yvon, 2005) show the application of analogy to morphological analysis in three different languages with results comparable or higher to that of another proposed technique; (Claveau and L'Homme, 2005) show how to guess the meaning of unknown terms using analogical proportions; (Lepage and Denoual, 2005b) use proportional analogies to translate sentences in a system that compares well with state-of-the-art machine translation systems; (Langlais and Patry, 2006) propose to specialize the previous technique to translate unknown words.

In linguistics, proportional analogy depreciated at the birth of historical linguistics when phonetic laws were considered the only scientific objects worth studying. In this respect, (Anttila, 1977) is a virulent defense of analogy with linguistic change as its main perspective. Still, until now, some linguists doubt the value of analogy because of the controversial argument of the poverty of the stimulus which claims that there is no inductive

device like analogy in language acquisition because (1) young children produce sentential structures they have never heard before and (2) they never produce some ungrammatical structures.[1] For instance, in the case of auxiliary fronting, young children can produce sentences like:

Is the student who is in the garden hungry?

the structure of which they have never heard before, and they never produce:

**Is the student who in the garden is hungry?*

although both sentences are formal solutions of the following analogical equation formed of sentences that they may have heard before:

$$\begin{array}{l}\textit{The student in the garden is hungry.}\end{array} : \begin{array}{l}\textit{Is the student in the garden hungry?}\end{array} :: \begin{array}{l}\textit{The student who is in the garden is hungry.}\end{array} : x$$

Against the argument, (Pullum and Scholtz, 2002) show that the structure in question actually appears in books for children and in the CHILDES corpus and support some of their claims with counting explanation. (Legate and Yang, 2002) however disagree with these countings and the debate is still open.

3 True proportional analogies

Of interest to the present article is a remark by (Pullum, 1999), that says that analogy cannot cross some structural boundaries. *E.g.*, the following analogy overlooks clause boundaries and does not make sense:

$$\begin{array}{l}\textit{white skirt}\end{array} : \begin{array}{l}\textit{green blouse}\end{array}$$

::

$$\begin{array}{l}\textit{Often commentators who are white skirt the problem of institutional racism.}\end{array} : \begin{array}{l}\textit{*Often commentators who are green blouse the problem of institutional racism.}\end{array}$$

Acknowledging this fact, the purpose of this article is to test whether analogies between chunks, *i.e.*, sequences of words delimited by such boundaries, appear in a sufficient number so as to give support and open the door to further use of analogy in natural language processing.

[1] Volume 9, 2002, of the Linguistic Review contains four pros and cons articles.

To be useful, analogies need to be 'true', *i.e.*, they have to be valid on the level of form *and* meaning. This is usually the case in conjugation where analogy explains paradigms:

to walk : *I walked* :: *to laugh* : *I laughed*

Conversely, there are analogies of form which are not analogies of meaning. Here is Chomsky's famous example:[2]

$$\begin{array}{l}\textit{Abby is baking vegan pies.}\end{array} : \begin{array}{l}\textit{Abby is baking.}\end{array} :: \begin{array}{l}\textit{Abby is too tasteful to pour gravy on vegan pies.}\end{array} : \begin{array}{l}\textit{Abby is too tasteful to pour gravy on.}\end{array}$$

And there are analogies of meaning which cannot be supported by an analogy of form:[3]

$$\textit{I swim} : \begin{array}{l}\textit{I'd like to swim}\end{array} :: \textit{I can swim} : \begin{array}{l}\textit{I'd like to be able to swim}\end{array}$$

4 Experimental protocol and data

In order to count 'true' analogies between chunks in a real corpus, we propose the following steps:

1. chunk the texts;
2. gather, by machine, all analogies of form between chunks;
3. sample the set of analogies of form;
4. filter, with the help of a human evaluator, the analogies of form contained in the sample according to their truth;
5. apply a statistical test to determine the proportion of true analogies on the collection of all analogies of form.

Each step can be performed automatically, except step 4 where human intervention is required.

The data we use are from the machine translation evaluation campaign IWSLT 2005 (Eck and Hori, 2005). They consist in 20,000 aligned short sentences in English and Japanese in the tourism domain. We used them as two different corpora, however keeping in mind the fact that they are comparable corpora. Some statistics on sizes in words and characters are to be found in Table 1.

[2] Noam Chomsky, *Conference at the university of Michigan*, 1998, a report by Aaron Stark. In the third sentence, gravy is put on the vegan pies whereas it is poured on Abby in the fourth sentence. Hence, the difference in structures in sentences 3 and 4 is not parallel to the one between 1 and 2.

[3] * *I'd like to can swim* is ungrammatical.

Table 1: Some statistics on the data. English nominal chunks are at least three word long. English verbal nuclei reduce to only one word. Japanese chunks were obtained with a list of markers automatically computed to deliver the highest number of analogies between chunks. The number of words in Japanese was measured using ChaSen.

Data type		Data size		
		number	in words	in characters
English	Sentences	20,000	188,935	842,722
	Nominal chunks	5,928	19,991	126,696
	Verbal nuclei	1,523	1,523	11,462
Japanese	Sentences	20,000	173,091	353,780
	Chunks (altogether)	99,719	693,526	718,819

5 Experiments

5.1 Chunking

Although there is no universally agreed definition of chunks, they are usually defined as elementary syntagmatic units. For instance, the following sentence in English may be divided into the following chunks:

$$[It][\textit{'s just}][\textit{down the hall}].$$

The English chunker we used is a command built on the top of TreeTagger.[4] For the previous sentence, the actual output is:

$$[It]_{NC}\ [\textit{'s}]_{NC}\ [\textit{just down}]_{ADVC}\ [\textit{the hall}]_{NC}.$$

As the chunks are categorized, it is possible to separate nominal chunks from verbal nuclei. However, the quality of the chunks delivered is very low (*e.g.*, in the example above, *'s* is categorized as a noun phrase maybe because of the Saxon genitive). We thus restricted ourselves to nominal chunks composed of three words or more, and to verbal nuclei reduced to only one word because those appeared to be categorized with a high degree of accuracy.

As for Japanese, there exists a standard chunker, YamCha,[5] but training data are needed to feed a training phase. As we did not have any such data at our disposal, we decided to adopt another standard approach to chunking in languages like Japanese. Indeed, Japanese is a language where cases are marked and, hence, it exhibits free constituent order. Markers constitute a closed set of words that appear at the end of nominal or verbal chunks. For instance, in the following Japanese sentence from our data (translation: *usually on business, seldomly for pleasure*), the words in uppercase are such case markers:

[*taitei shigoto DE*] [*metta NI*] [*asobi DE WA*]
[*often FOR work*] [*seldomLY*] [*FOR pleasure*]

To determine the markers, we used a hill-climbing method that automatically selects the most productive ones relatively to the number of analogies between the chunks. In Japanese, no blanks seperate words in usual texts and it is worth noticing that we did not segment into words before chunking.[6]

It was of course tempting to check whether the same method (in characters and in markers) could be applied onto the English data. We selected the shortest most frequent sequences in the English data and used them as markers to cut down into units that we expected to be chunks.[7] However, the results were too poor to be exploitable.

To summarize, our chunking methods are highly language-dependent, and they reflect to some extent the basic properties of the two languages at hand: the fixed word order of English implies that position is the main clue for syntactic functions whereas Japanese indicates funtions by case markers, and is thus a language with free consituent order.

The number of chunks obtained are very different in the two languages. In English we retained only 6,000 nominal chunks of three

[4] Institute for Computational Linguistics of the University of Stuttgart, `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[5] Nara Advanced Institute of Technology, (Kudo and Matsumoto, 2003)

[6] The standard tool for Japanese segmentation, ChaSen, available with any Linux distribution, is also from Nara Advanced Institute of Technology.

[7] Apart from punctuation marks, the list comprised, of course, such words as: *the, a, is, to, ...*

words or more and 1,500 verbal nuclei of one word, while in Japanese, we obtained almost 100,000 chunks of any type.

5.2 Gathering analogies of form

The formalization of proportional analogies of form we adopted here follows the proposal in (Lepage and Denoual, 2005b).[8] It reduces to the counting of number of symbol occurrences and the computation of edit distances between strings of characters:

$A : B :: C : D \quad \Rightarrow$

$$\begin{cases} |A|_a + |D|_a &= |B|_a + |C|_a, \forall a \\ \text{dist}(A,B) &= \text{dist}(A,C) \\ \text{dist}(A,C) &= \text{dist}(A,B) \end{cases}$$

where $|A|_a$ stands for the number of occurrences of character a in string A and $\text{dist}(A,B)$ is the edit distance between strings A and B. To extract all possible analogies of form between chunks, we thus consider chunks as ordinary strings of characters whatever the character set, the Latin alphabet or the Japanese kanji-kana character set.

The previous formalization permits to avoid a naive approach of the computation of all possible analogies that would be in $O(N^4)$ for N character strings. The sparseness of the search space is exploited by first looking for those 4-tuples (A,B,C,D) such that $|A|_a - |B|_a = |C|_a - |D|_a$. This is tantamount to look for the set of pairs (A,B) such that $|A|_a - |B|_a = n_a$ for all possible values of vectors (n_a) where a scans the character set. By sorting all vectors in lexicographic order and in decreasing order of the numerical values, one may incrementally inspect relevant pairs only. For these relevant pairs with the same vector value, one can, in last instance, evaluate the truth of $\text{dist}(A,B) = \text{dist}(C,D)$. The use of bit representations techniques yields tractable computational times. On a 2.2 GHz processor we were able to gather all analogies on the English nominal chunks and verbal nuclei in some hours and we needed two days for the nearly hundred thousand Japanese chunks (the size of the search space for this latter case being theoretically of 10^{20}!).

[8] Rather than the more complex form of (Delhay and Miclet, 2004) or another proposal in terms of automata (Stroppa and Yvon, 2005).

The different numbers of analogies of form between chunks are given in Table 2. For English, the number of analogies of form between nominal chunks is less than ten thousand for nearly 6,000 chunks, whereas the number of analogies slightly exceeds 40,000 for only 1,500 English verbs. For Japanese, the figure exceeds 2 million for nearly hundred thousand chunks. We observed that the production of analogies of form in Japanese is much higher than that for English nominal chunks, and that the number of analogies between English verbs is proportionally the highest one.

5.3 Sampling

For each category of chunks obtained (English nominal chunks, English verbal nuclei and Japanese chunks), we drew a sample of 100 analogies from the set of analogies automatically gathered as explained in the previous section. Although this is statistically unnecessary, we indeed repeated the sampling, filtering and testing steps a few times and confirmed the results reported below.

5.4 Filtering

Each analogy in the sample was presented to an annotator whose task was to estimate its validity in meaning so as to establish its truth, *i.e.*, its validity in form and meaning. This task was carried out using a browser interface. Each analogy was presented to the annotator one after another and the annotator had to check a radio box to invalidate an analogy as being a true analogy before going to the next analogy of form. At the end of the task, a summary presented the annotator with a number of pieces of information: the percentage of true analogies, the p-value for the null-hypothesis, 5 examples of true analogies and 5 examples (if possible) of analogies of form that were not considered analogies on the level of meaning (see Figure 1).

5.5 Testing

As there are only two issues in this experiment – an analogy may be true or false – we applied a binomial test to test a null hypothesis of 96% of the analogies being true analogies. The figure of 96% comes from (Lepage and Denoual, 2005b) who reported it for a collection of 160,000 short English sentences (p-value of 0.1%).

Table 2: Number of analogies of form and estimation of the number of true analogies, *i.e.*, analogies of form and meaning with a 96% null hypothesis.

Data type		Number of chunks	Number of analogies of form	Number of true analogies	
				Observed percentage and p-value	Inferred absolute number
English	Nominal chunks	5,928	19,991	98% (0.001)	19,591
	Verbal nuclei	1,523	40,525	96% (0.002)	38,904
Japanese	Chunks (altogether)	99,719	2,131,269	96% (0.005)	2,046,018

Table 2 summarizes the results of the tests for each type of chunks. The null hypothesis of 96% true analogies has been verified for all the types of chunks. Only a few analogies of form have been judged invalid in meaning.

In absolute and raw figures, this estimate yields:

- almost 20,000 true analogies for nearly 6,000 English nominal chunks;
- nearly 39,000 true analogies for thousand and a half English verbs;
- two million true analogies for almost hundred thousand Japanese chunks.

6 Synthesis of the results and further research

Two points are worth further inspection in combination with other results reported in the litterature: (1) the ratio of true analogies to the number of chunks and (2) the proportion of true analogies relatively to analogies of form in English.

The ratio of true analogies to the number of chunks is the average number of times a chunk enters a true analogy. As analogies reflect commutation operations between and within several pieces of language, this ratio measures the 'combinatorial structural power' of chunks in the corpus considered. For each different type of chunks, we report the following figures:

- 3.3 for English nominal chunks (19,591/5,928);
- 25.5 for English verbs (38,904/1,523);
- and 20.5 for Japanese chunks (2,046,018/99,719).

Within the frame of this paper, we are left to observe that the figures obtained are very different. Further research in complement to the experiments reported here should graph the increase of true analogies with the number of chunks, so as to set out experimental laws of 'combinatorial structural power' for each of the types of chunks. Such laws should indeed be determined for any piece of language, like words and sentences.

As for the proportion of true analogies relatively to analogies of form for English only, a rough figure of at least 95% summarizes several experimental results as this figure has been shown to hold for:

- verbal forms (one word, this paper);
- nominal chunks of three words or more (this paper);
- short sentences of around seven words (Lepage and Denoual, 2005b).

This figure of 95% somehow relates to *competence*, because it establishes the truth of analogies between words, chunks or sentences attested in their respective corpora. But it should also be put in connection with a result given by (Lepage and Denoual, 2005a) that relates to *performance*. The authors of this paper report experiments in producing paraphrases that were shown to be semantically correct in 94% of the cases, either by equivalence or entailment, with some specific analogical method. Further studies on the performance side and in different languages, like Japanese for instance, should be conducted to verify or falsify this figure of around 95%.

$$\textit{a dollar bill} : \textit{five thousand yen} :: \textit{a hundred dollar bill} : \textit{five hundred thousand yen}$$

$$\textit{any free rooms} : \textit{some free time} :: \textit{any other rooms} : \textit{some other time}$$

$$\begin{array}{l}\textit{my insurance}\\ \textit{company}\end{array} : \begin{array}{l}\textit{your insurance}\\ \textit{company}\end{array} :: \begin{array}{l}\textit{my passport}\\ \textit{and ticket}\end{array} : \begin{array}{l}\textit{your passport}\\ \textit{and ticket}\end{array}$$

$$\begin{array}{l}\textit{any concert tonight}\\ \textit{and tomorrow night}\end{array} : \begin{array}{l}\textit{any concerts tonight}\\ \textit{and tomorrow night}\end{array} \not:: \textit{this Mr. Ono} : \textit{this Mrs. Ono}$$

Figure 1: Three examples of true analogies bteween English nominal chunks and one analogy of form that is not an analogy of meaning (noted by $\not::$).

7 Conclusion

In this paper, we measured the number of true proportional analogies between chunks in two typologically different languages on a 20,000 sentence long Japanese-English bicorpus. We verified that at least 96% of analogies of form between chunks are also analogies of meaning. This shows that analogy is a reliable structuring device between chunks.

Still, the absolute number of analogies collected in the two languages are quite different due to the fact that our chunking techniques somehow reflected the typologically different nature of the two languages. As an open question, one may ask whether maximizing the number of analogies between chunks could help to improve chunking, whatever the language.

References

Raimo Anttila. 1977. *Analogy*. Mouton – Trends in linguistics: state of the art reports 10, The Hague.

Vincent Claveau and Marie-Claude L'Homme. 2005. Terminology by analogy-based machine learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. Copenhagen (Denmark).

Arnaud Delhay and Laurent Miclet. 2004. Analogical equations in sequences: Definition and resolution. *Lecture Notes in Computer Science*, 3264:127–138.

Thomas Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In Carnegie Mellon University, editor, *Proc. of the International Workshop on Spoken Language Translation*, pages 1–22.

Tadu Kudo and Yuji Matsumoto. 2003. Fast methods for kernel-based text analysis. In *Proceedings of ACL 2003*, pages ??–??, ??, July.

Philippe Langlais and Alexandre Patry. 2006. Translating unknown words by analogical learning. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 877–886.

Julie Anne Legate and Charles D. Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19:151–162.

Yves Lepage and Etienne Denoual. 2005a. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *Proceedings of the third international workshop on Paraphrasing (IWP 2005)*, pages 57–64, Jeju, October.

Yves Lepage and Etienne Denoual. 2005b. Purest ever example-based machine translation: detailed presentation and assessment. *Machine Translation Journal*, 19:251–282.

Geoffrey K. Pullum and Barbara C. Scholtz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19:9–50.

Geoffrey K. Pullum, 1999. *Generative grammar*, pages 340–343. The MIT Press, Cambridge.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pages 120–127, Ann Arbor, MI, June.

Peter D. Turney and M.L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3):251–278.