Analogies of form between chunks in japanese are massive and far from being misleading

Yves Lepage, Julien Migeot, Erwan Guillerm

Analogies of form between chunks in Japanese are massive and far from being misleading

Yves Lepage, Julien Migeot and Erwan Guillerm

GREYC, University of Caen, BP 5186, Caen Cedex, France
Yves.Lepage@info.unicaen.fr, {Erwan.Guillerm,Julien.Migeot}@etu.info.unicaen.fr

Abstract

This paper relates to the assessment of the argument of the poverty of the stimulus in that we measure the number of true proportional analogies between chunks in a language with case markers, Japanese. On a bicorpus of 20,000 sentences, we show that at least 96% of the analogies of form between chunks are also analogies of meaning, thus reporting the presence of at least two million true analogies between chunks in this corpus. As the number of analogies between chunks grows nearly quadratically compared to sentences, we conclude that proportional analogy is an efficient and undeniable structuring device between chunks.

1. Introduction

The purpose of this paper is to make an experimental contribution to the discussion of the argument of the poverty of the stimulus, by inspecting the reality of true analogies in syntax[1]. Hence, we shall question the reality, estimate the amount and assess the truth of analogies between chunks. This obviously relates to the usefulness of analogy in terms of linguistic performance, and in terms of language acquisition.

The paper is organised as follows. Section 2 briefly sets the scene for the argument of the poverty of the stimulus to justify the purpose of the measure reported here. Section 3 illustrates the notion of true analogy by examples. Section 4 describes the experimental protocol used and gives details about a formal definition of analogy, the method used for chunking and, statistical tests. The results are summarized and analyzed in Sections 5 and 6.

2. The argument of the poverty of the stimulus

The argument of the poverty of the stimulus is a controversial argument in the study of language and mind (see volume 19 of the Linguistic Review: (PULLUM and SCHOLTZ, 2002), (LEGATE and YANG, 2002), (SCHOLTZ and PULLUM, 2002)). It assumes that the information in the environment would not be rich enough to allow a human learner to attain adult competence in his/her native language. More precisely, the argument is based on the controversed fact that young children would produce some sentential structures they would have never heard before. In addition, according to the proponents of the argument of the poverty of the stimulus, if some sentential structures would be derived by an induction device like analogy, then, children would also derive ungrammatical structures which, accordingly to the proponents of this argument, they never utter.

A representative example of such a structure is auxiliary fronting in interrogative sentences that involve a relative clause. Positing analogy as the induction device in language acquisition would imply that children would indifferently produce such sentences as:

Is the student who is in the garden hungry?
and **Is the student who in the garden is hungry?*
because both sentences are valid formal solutions to the following analogical equation built from sentences that they may well have heard before:

The student in the garden is hungry.	:	*Is the student in the garden hungry?*	:: *The student who is in the garden is hungry.* : x

(PULLUM and SCHOLTZ, 2002) objected the hypothesis that children would have never heard the structure in question, by showing that it does appear in books for children and in the CHILDES corpus. Also, the argument partly relies on the assumption that children would learn exclusively by positive examples, an hypothesis objected in (CHOUINARD and CLARK, 2003), where it is shown that children between the age of two and three and a half do produce ungrammatical sentences, that adults do correct them, and that children do repeat the corrected utterances in 50 to 60% of the cases, which indicates that children understand correction, and consequently, that they can memorize pairs of negative and positive examples.

A defender of analogy, (PULLUM, 1999) acknowledges the fact that analogy cannot overlook some grammatical boundaries without the risk of producing meaningless utterances, as in:

white skirt	:	*green blouse*
Often commentators who are white skirt the problem of institutional racism.	::	**Often commentators who are green blouse the problem of institutional racism.*

[1] What is addressed here is different from the view that puts analogy in opposition with phonetic change and language change in general, see (ANTTILA, 1977).

3. Goal of the paper and true analogies

The purpose of this paper is to give support to the proponents of analogy in syntax, by testing the reality of proportional analogies between the most elementary grammatical units, *i.e.*, chunks. The way we achieved this is by gathering all possible analogies of form between chunks extracted from a corpus and by estimating the number of true analogies.

True analogies are proportional analogies which are valid on the level of form and that are meaningful. They are best illustrated in declension and conjugation where they explain paradigms. For instance:

$$\textit{to walk} : \textit{I walked} :: \textit{to laugh} : \textit{I laughed}$$

Conversely, misleading analogies of form which are not analogies of meaning have been illustrated by Chomsky's famous example in syntax[2]:

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

Analogies of meaning which are not supported by an analogy of form are illustrated below[3]:

$$\textit{I drink} : \textit{I'd like to drink} :: \textit{I can swim} : \textit{I'd like to be able to swim}$$

4. Experimental protocol

As a logical consequence of what has been said above, a counting of true analogies between chunks on a real corpus can adopt the following steps where all steps can be performed automatically, except for Step 4. where human intervention is required.

1. chunk the texts;
2. gather, by machine, all analogies of form;
3. sample the set of analogies of form if it is too big;
4. filter, with the help of a human annotator, the analogies of form contained in the sample that are true analogies;
5. apply a statistical test to determine the proportion of true analogies on the collection of all analogies of form.

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

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

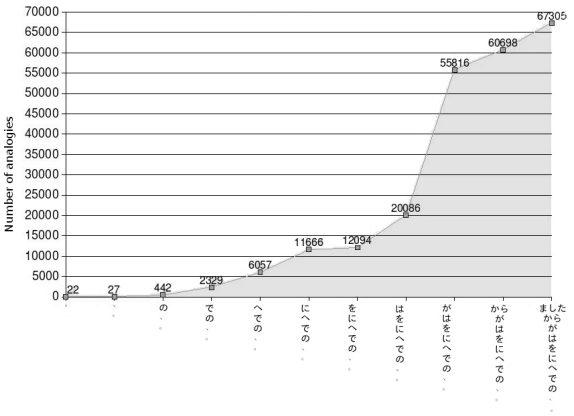

Figure 1: Number of analogies between chunks obtained on a 3,500 line sample of our corpus with different markers for chunking. One chunk marker is added to the list of markers at each step. The order of addition is determined by hill-climbing. Hence, at each step, the increase in number of analogies is the highest possible. Of all the markers, *ga* (subject), added at step 9, yields the most important increase.

The data used are from the machine evaluation campaign IWSLT 2005 (ECK and HORI, 2005). They consist in 20,000 sentences in Japanese. Some statistics on sizes in characters and words are to be found in Table 1.

4.1. Step 1. Chunking

Chunking is the process by which a sentence is divided into chunks. There exists a standard chunker for Japanese, YamCha (KUDO and MATSUMOTO, 2003), but training data are required in order to feed a training phase. First, we did not have any extra data at our disposal, and, second, part of our interest in this study resided in testing another standard approach to chunking for languages like Japanese. Indeed, Japanese is a language with cases markers, a closed set of words (or morphemes) appearing at the end of chunks. For instance, in the following transcribed Japanese excerpt from our data (translation: *usually on business, seldom 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 finally used, we started with 16 well identified nominal case markers (9), verbal endings (5) and punctuations (2). The most productive ones in terms of number of analogies were automatically selected using a hill-climbing method on a sampling of 3,500 lines from our data. The 11 most productive markers for which we observed significant increase until a plateau

are, in this order, the two punctuation marks, 8 nominal case markers and only one verbal ending[4]. The rate of increase in number of analogies for the retained 11 markers is shown on Figure 1.

4.2. Step 2. Gathering analogies of form

The next problem is to extract all possible analogies between the elements of a corpus, be the elements sentences or chunks. From the program point of view, the elements are just strings of characters, whatever the character set, the Latin alphabet or the Japanese kanji-kana character set.

To leave out trivial cases of analogies, the implemented program inspects only analogies of the type $A : B :: C : D$ where the character strings A, B, C and D are all different.

The formalisation of proportional analogies of form adopted here follows the proposal in (LEPAGE, 2004)[5]. From the programming point of view, it reduces to the computation of edit distances and the counting of number of symbol occurrences. Precisely:

$$A : B :: C : D \Rightarrow \begin{cases} \mathrm{dist}(A,B) = \mathrm{dist}(A,C) \\ \mathrm{dist}(A,C) = \mathrm{dist}(A,B) \\ |A|_a + |D|_a = |B|_a + |C|_a, \quad \forall a \end{cases}$$

where $\mathrm{dist}(A, B)$ is the edit distance between strings A and B and $|A|_a$ stands for the number of occurrences of character a in string A.

A naïve approach of the computation of all possible analogies between the N elements of a corpus would examine all possible 4-tuples and would thus be in $O(N^4)$, an asymptotic behaviour that is simple unaffordable for the size of the corpus we work with: nearly hundred thousand chunks.

The formalization allows to exploit the sparseness of the search space by first looking for those 4-tuples (A, B, C, D) such that $|A|_a - |B|_a = |C|_a - |D|_a$, as it is tantamount to look inside different sets 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 the 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 $\mathrm{dist}(A, B) = \mathrm{dist}(C, D)$.

[4]The punctuations are the symbols corresponding to fullstop and comma. The next nominal case markers are *no* (genitive), *de* (instrumental or location), *e* (direction), *ni* (dative or location), *wo* (accusative, *i.e.*, object), *wa* (topic), *ga* (subject), *kara* (origin). The verbal ending is, surprisingly, the past ending *-masita*.

[5]Rather than the more complex form of (DELHAY and MICLET, 2004) or another proposal in terms of automata (STROPPA and YVON, 2005).

The use of bit representations techniques, even for distance computation (ALLISON and DIX, 1986), yields tractable computational times. We were able to gather all analogies from nearly hundred thousand chunks in two days on a 2,2 GHz processor (the size of the search space being theoretically of 10^{20}!) and in less than ten minutes for the set of chunks extracted from 3,500 sentences (approximate time for each step in Figure 1).

4.3. Steps 3., 4. and 5. Sampling, filtering and testing

As the number of analogies of form automatically gathered is untractable by hand, we had to sample them. Each analogy in the sample was then presented to an annotator whose task was to estimate its validity in meaning so as to establish the truth of the analogy, *i.e.*, its validity in form and meaning.

This task was carried out using a browser interface. Each analogy is presented to the annotator one after another and the annotator has 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 presents the annotator with a number of pieces of information: 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 analogy on the level of meaning.

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. This figure of 96% comes from (LEPAGE, 2004) who reported it for a collection of 160,000 short Chinese, English and Japanese sentences.

5. Results

5.1. Chunking

Coming back to chunking (Step 1.), the increase in number of chunks that we observed is almost perfectly linear in the number of lines (see table below). Such a result meets intuition as it simply means that the data are well distributed and consist in lines of 5 chunks in average.

Number of lines	Number of chunks
500	2462
1000	5070
1500	7531
2000	9909
2500	12316
3000	15042
3500	17357
5000	25121
10000	50117
15000	74684
20000	99719

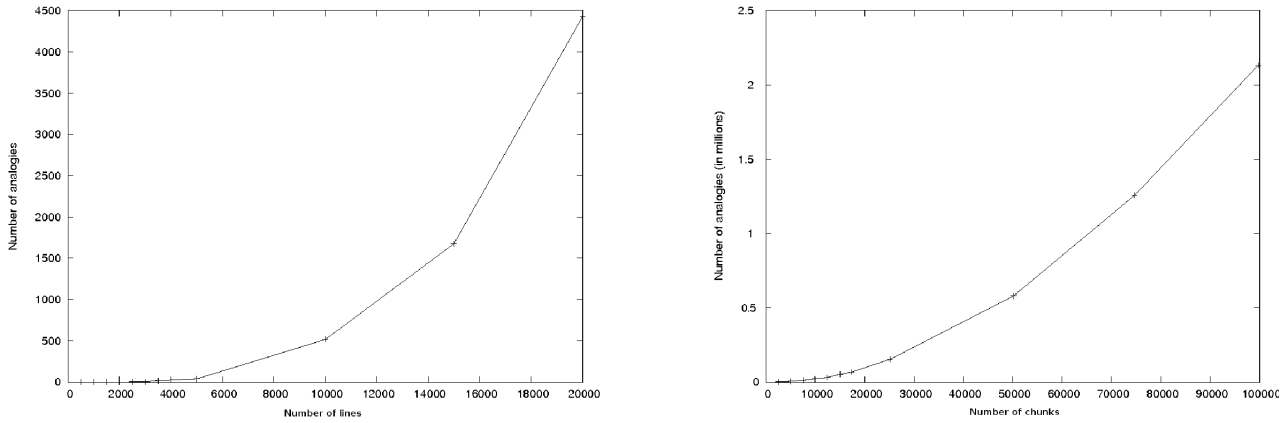

Figure 2: Number of analogies against the number of sentences (on the left) or chunks (on the right). Caution: the ordinate scale is a different order of magnitude in both graphs.

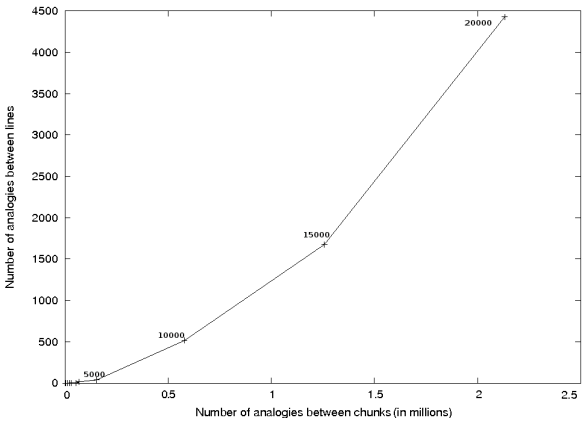

Figure 3: Number of analogies bteween chunks in abscissae, number of analogies between lines (sentences) in ordinates.

5.2. Counting analogies of form

As for gathering analogies of form (Step 2.), the graph on the left of Figure 2 plots the number of analogies between sentences against the number of lines (sentences). Until 2,500 lines, no analogies are found. After this value, the increase looks at least polynomial.

On the right of the figure, the number of analogies between chunks has been plotted against the number of chunks (the chunks being gradually obtained from the lines). Notice that the ordinate scales of the two graphs in Figure 2 differ by an order of 3 digits: 4,428 analogies only for 20,000 lines; 2,131,269 analogies for the 99,719 chunks obtained from these very 20,000 lines.

The graph of Figure 3 plots the number of analogies between chunks in abscissae, against the number of analogies between lines (sentences) in ordinates (the chunks coming only from the corresponding sentences). The increase looks polynomial and we could make the experimental number of analogies between chunks fit with the number of analogies between sentences taken to the power of 1.7 (Pearson correlation: 0.99, all other powers yielding a lesser correlation). The relationship can thus be considered nearly quadratic.

5.3. Estimating the percentage of true analogies

The two final figures, that were the actual goal of this study, are the percentage of true analogies observed on a few samples of 100 analogies. They are given on the right of Table 1.

On this corpus of 20,000 Japanese sentences, it is estimated that all analogies of form gathered between sentences are analogies of meaning, making them all true analogies. This important result can be interpreted as follows: the kind of examples quoted in Section 3. (Abby and the gravy) that indeed look artificial, may in fact happen very scarcely in real utterances in comparison with analogies of form that are actually analogies of meaning.

Table 1: Statistics for the data and estimation of the number of true analogies, *i.e.*, analogies of form and meaning with a 96% null hypothesis. ChaSen was used to count Japanese words. On average, a chunk appears 4.1 times in the corpus.

Data type	Data size			Average size	Number of analogies of form	Number of true analogies	
	in lines	in words	in characters	in words		% observed	p-value
Sentences	20,000	173,091	339,579	8.7	4,428	100%	n.r.
Chunks	99,719	693,526	718,819	6.9	2,131,269	96%	0.005

As for chunks, the null hypothesis of 96% true analogies has been verified. Only a few analogies of form have been judged invalid in meaning. As has already been mentioned, the number of analogies of form between chunks is enormous in comparison with the number of sentences, and the result of 96% true analogies should be considered bearing this explosion in mind. In absolute figures, our estimate of at least 96% of the analogies of form being true analogies yields an absolute figure of at least two million true analogies between chunks (precisely: 2,131,269 $\times$ 96% = 2,046,018) for nearly hundred thousand chunks (precisely: 99,719). A possible interpretation is that, in average, each chunk takes part in 20 true analogies.

6. Analysis of the results and conclusion

In this paper, we addressed the problem of the reality, the amount and the truth of analogies between chunks contained in a Japanese corpus.

The amount of analogies gathered and the estimation of the number of true analogies obtained, *i.e.*, analogies of form and meaning, establish in an undeniable manner the reality of proportional analogies.

We obtained more than two million analogies of form between chunks extracted from a corpus of 20,000 short sentences, each sentence containing an average of five chunks. As for their truth, we estimated that more than 96% of the analogies of form are true analogies. These figures are in blatant contradiction with the opinion that analogies of form would almost necessarily lead to nonsense and would have weak connection with meaning.

The results obtained here are promising because they show that analogies can be exploited in natural language processing applications at a higher level than the ordinary level of words, as in (STROPPA and YVON, 2005), or terms, as in (CLAVEAU and L'HOMME, 2005). On the other hand, our comparison of sentences with chunks has shown that, on our data, the number of analogies between chunks grows as nearly the square of the number of analogies between those sentences containing these chunks. An experimental conclusion that can be drawn is that chunks may be the most productive level of application for true proportional analogies.

7. References

ALLISON, Lloyd and Trevor I. DIX, 1986. A bit string longest common subsequence algorithm. *Information Processing Letter*, 23:305–310.

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

CHOUINARD, Michelle M. and Eve E. CLARK, 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30:637–669.

CLAVEAU, Vincent 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).

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

ECK, Thomas and Chiori HORI, 2005. Overview of the IWSLT 2005 evaluation campaign. In Carnegie Mellon University (ed.), *Proc. of the International Workshop on Spoken Language Translation*.

KUDO, Tadu and Yuji MATSUMOTO, 2003. Fast methods for kernel-based text analysis. In *Proceedings of ACL 2003*. ??

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

LEPAGE, Yves, 2004. Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus. In *Proceedings of COLING-2004*, volume 1. Genève.

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

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

SCHOLTZ, Barbara C. and Geoffrey K. PULLUM, 2002. Searching for arguments to support linguistic nativism. *The Linguistic Review*, 19:185–224.

STROPPA, Nicolas and François YVON, 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*. Ann Arbor, MI.