A study of the saturation of analogical grids agnostically extracted from texts

Rashel Fam and Yves Lepage *

IPS, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
fam.rashel@fuji.waseda.jp, yves.lepage@waseda.jp

Abstract. Analogical grids aim to capture the organization of the lexicon of a language. We conduct experiments on analogical grids extracted in four different languages with different morphological richness. We study the saturation of analogical grids against their size. We observe that the logarithm of the saturation of an analogical grid is linear in the logarithm of its size. More surprisingly, the coefficients of this log-log linear relation are extremely close across all four languages, even when the size or the genre of the corpus vary.

Keywords: analogical grids, saturation, organization of lexicon.

1 Introduction and background

show	*shows*	*showing*	*showed*
walk	*walks*	*walking*	*walked*
open	*opens*	*opening*	
study		*studying*	
read	*reads*	*reading*	

makan	*dimakan*	*memakan*	*makanan*
minum	*diminum*	*meminum*	*minuman*
main			*mainan*
beli	*dibeli*		

Fig. 1. Analogical grids in English (*left*) and Indonesian (*right*).

Figure 1 shows two examples of analogical grids, one in English, the other one in Indonesian. Such analogical grids may be automatically constructed from the set of words contained in a text. Each cell in an analogical grid either contains a word form or is empty. As exemplified in Figure 1 (*left*), a column (or a row) in an analogical grid usually exhibits similar word forms for different words: e.g., infinitive, present 3rd person singular, present participle, etc. for different English verbs on the *left* of Figure 1. Analogical grids are not paradigm tables,

* This work was supported by a JSPS Grant, Number 15K00317 (Kakenhi C), entitled Language productivity: efficient extraction of productive analogical clusters and their evaluation using statistical machine translation.

i.e., they are not the result of a linguistic formalization with explicit lexemes and exponents as in standard works in morphology, but they constitute a preliminary step in that direction. Analogical grids too give a compact view of the organization of the lexicon, but they are the output of an empirical procedure, e.g., the one introduced in [4].

Analogical grids can be used to study word productivity in a given language as in [12, 9, 6]. They can also be used to make comparisons across languages as in [4], where the goal is to explain unseen words by using analogical grids automatically built from the set of all words contained in texts in 12 different languages.

In this paper, we report an interesting phenomenon observed when building analogical grids in various different languages using the method in [4]. This phenomenon relates the saturation of the obtained analogical grids to their size. The experimental results show that the coefficients which characterize the relation would not be influenced by the size, the genre or the language of the texts used.

The paper is organized as follows: Section 2 introduces basic notions related to analogical grids. Section 3 presents our experiments on four languages with different richness in morphology. It analyzes the results and explores the relationship between the saturation and the size of analogical grids. Section 4 presents further experiments to inquire the relation. Section 5 gives conclusion.

2 Basic notions

In this section, we mathematically define the basic notions related to analogical grids. The method to extract such analogical grids has already been presented elsewhere [8, 4].

2.1 Illustration with toy data

Anto memakan nasi dan meminum air. Nasi itu dibeli di pasar. Di pasar, Anto melihat mainan. Anto senang main bola. Setelah main, Anto suka minum es dan makan cilok. Makanan dan minuman itu juga dia beli di pasar. Es dan cilok memang enak dimakan dan diminum selesai olahraga.

air anto ***beli*** *bola cilok dan di dia* ***dibeli dimakan diminum*** *enak es itu juga* ***main mainan makan makanan*** *melihat* ***memakan*** *memang* ***meminum minum minuman*** *nasi olahraga pasar selesai senang setelah suka*

Fig. 2. A text in Indonesian (*above*) and the list of words extracted from it (*below*). Words appearing in Figure 1 (*right*) are boldfaced.

The top of Figure 2 is a forged example text in Indonesian, a language which is known for its relative richness in derivational morphology. We intentionally do

not give its translation into English to place the reader in the agnostic position of the computer in front of such data. The list of words, sorted in lexicographic order, that can be extracted from this text, is given at the bottom of Figure 2.

From this word list, some commonalities between words can be identified at a glance. An example is the word *makan* and the word *makanan*. Another is the words *bola* and *beli* which share the same consonants in the same order: *b* and *l*. However, the existence of only one pair is not enough to support the evidence that two words are actually in relation one with the other. On the contrary, for the words *makan* and the word *makanan*, the same *ratio* is seen to hold between several other word pairs from the same text, like *minum* and *minuman*, or *main* and *mainan*. These actually reflect a phenomenon in Indonesian morphology by using the suffix *-an* which builds a noun from active verb.

In standard linguistics, a systematization of these relationships between word forms is given by paradigm tables, which is the result of linguistic formalisation. Here, we agnostically extract analogical grids relying on a formal relationship between words, proportional analogy. The right part of Figure 1 shows the analogical grid extracted from the set of words given in Figure 2.

2.2 Analogical grids

An analogical grid is a table of dimension $M \times N$ as defined by Formula (1). As illustrated by Figure 1, analogical grids extracted from texts usually contain empty cells. (Caution: there is no importance in the order of lines or rows.)

$$\begin{array}{ccc} P_1^1 : P_1^2 : \cdots : P_1^m \\ P_2^1 : P_2^2 : \cdots : P_2^m \\ \vdots \quad \vdots \qquad \vdots \\ P_n^1 : P_n^2 : \cdots : P_n^m \end{array} \quad \overset{\Delta}{\Longleftrightarrow} \quad \begin{array}{r} \forall (i,k) \in \{1,\ldots,n\}^2, \\ \forall (j,l) \in \{1,\ldots,m\}^2, \\ P_i^j : P_i^l :: P_k^j : P_k^l \end{array} \tag{1}$$

The definition of analogical grids in Formula (1) implies that for any four word forms at the intersection of two rows and two columns form a proportional analogy between sequences of characters [7, 13]. A proportional analogy is defined as a relationship between four objects where two properties are met:

(a) equality of ratios (defined hereafter) between the first and the second terms on one hand, and the third and the fourth terms on the other hand, and
(b) exchange of the means (the second and the third terms can always be exchanged).

$$A : B :: C : D \quad \overset{\Delta}{\Longleftrightarrow} \begin{cases} A : B = C : D \\ A : C = B : D \end{cases} \tag{2}$$

According to Formula (1), we can get many analogies from analogical grids in Figure 1. Figure 3 shows three of them.

We define the ratio between two words in Formula (3) as a vector of features made up of all the differences in number of occurrences in the two words, for all

$$makan : makanan :: main : mainan$$
$$makan : memakan :: minum : meminum$$
$$minum : diminum :: beli : dibeli$$

Fig. 3. Some analogies extracted from analogical grid in Figure 1 (*right*).

the characters, whatever the writing system, plus, the distance between the two words.

$$A : B \triangleq \begin{pmatrix} |A|_a - |B|_a \\ |A|_b - |B|_b \\ \vdots \\ |A|_z - |B|_z \\ d(A,B) \end{pmatrix} \qquad makan : makanan \triangleq \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 2 \end{pmatrix} \tag{3}$$

In Formula (3), the notation $|S|_c$ stands for the number of occurrences of character c in string S. The last dimension, written as $d(A, B)$, is the edit distance between the two strings. This indirectly gives the number of common characters appearing in the same order in A and B.[1]

The above definition of ratios captures prefixing and suffixing. Although we do not show it here, this definition also captures parallel infixing or interdigitation, well-known phenomena in semitic languages [1, 14]. However, reduplication or repetition (e.g. consonant spreading) are not captured by this definition.

$$\underset{makan : makanan}{\begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 2 \end{pmatrix}} = \underset{main : mainan}{\begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 2 \end{pmatrix}} \quad \& \quad \underset{makan : main}{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 3 \end{pmatrix}} = \underset{makanan : mainan}{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 3 \end{pmatrix}}$$

$$\Rightarrow$$

$$makan : makanan :: main : mainan$$

Fig. 4. The two ratios between pairs of words for the first analogy in Figure 3.

This formal definition of word ratio in Formula (3) gives the same vector for the ratios $makan : makanan$, $makan : namakan$, and $makan : mnaakan$. This is due to the use of insertion and deletion as the only edit operations.

The purpose of working with analogical grid, and not only with individual analogies, is that Formula (1) imposes more constraints for a word form to enter

[1] The only two edit operations used are insertion and deletion, hence, $d(A, B) = |A| + |B| - 2 \times s(A, B)$. $|S|$ denotes the length of a string S and $s(A, B)$ is the length of the longest common sub-sequence (LCS) between A and B.

a grid: a word form in a grid must satisfy all analogy relationship with all surrounding word forms in the grid. The word form *makanan* in the analogical grid of Figure 1 (right) is the only word form which fits in, among *makanan*, *namakan*, or *mnaakan*. For example, as proved below, using the words *main* and *mainan* from the analogical grid, the inequality between the ratios *makan* : *main* and *namakan* : *mainan* implies that there is no analogy between these four words. The same holds for the word form *mnaakan*. In all these cases, the inequality comes from different edit distance values.

$$\begin{array}{ccc} makan : main & & namakan : mainan \\ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 3 \end{pmatrix} & \neq & \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 5 \end{pmatrix} \end{array} \quad \Rightarrow makan : main \not:: namakan : mainan$$

The above discussion shows that there should be a relationship between the size of the analogical grids and the freedom in filling an empty cell in an analogical grid.

2.3 Size and saturation of analogical grids

We simply define the size of an analogical grid as its number of rows multiplied by its number of columns. The analogical grids in Figure 1 has a size of $4 \times 5 = 20$ (*left*) and $4 \times 4 = 16$ (*right*) respectively.

Let us now turn to the number of empty cells of an analogical grid, or rather the number of non-empty cells which we call its *saturation*[2]. We compute it using Formula (4) which will give a saturation of 80 % (*left*) and 75 % (*right*) for Figure 1.

$$\text{Saturation} = 100 - \frac{\text{Number of empty cells} \times 100}{\text{Total number of cells}} \tag{4}$$

3 Experiments

3.1 Data used

We carried out experiments on a multilingual parallel corpus created from the translation of the Bible collected by Christodoulopoulos[3] [10]. We selected four languages with different richness in morphology: English, Russian, Modern Greek, and Indonesian. The reason for using a multilingual parallel corpus is the need to draw conclusions across different languages in a reliable way. Table 1 presents statistics on the corpus. For each text in each language, we first extracted the list of all words, and finally built all analogical grids.

[2] In [2, p. 79], saturation is the maximal proportion of word forms attested for any one lemma of a given paradigm. Here we use the term for each entire grid.

[3] http://homepages.inf.ed.ac.uk/s0787820/bible/

Language	# tokens (N)	# types (V)	Length of types avg±std. dev.	# grids	Time (h:min)
English	792,074	12,498	7.03 ± 2.18	12,855	45
Indonesian	648,606	15,641	7.84 ± 2.63	25,752	2:04
Modern Greek	706,771	36,786	8.49 ± 2.49	69,173	11:03
Russian	560,524	47,226	8.26 ± 2.73	60,035	10:34

Table 1. Statistics on the Bible corpus and number of analogical clusters and number of analogical grids produced in each language with the time needed to produce them

3.2 Analogical grids obtained

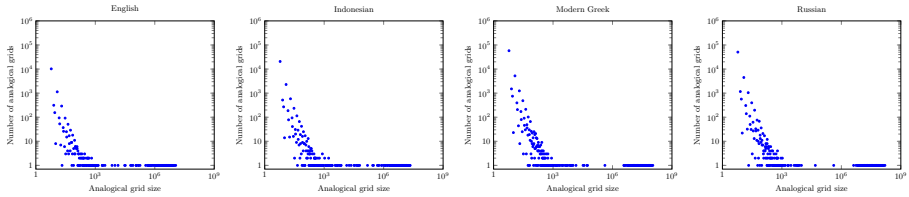

Fig. 5. Number of analogical grids with the same size in each language. Logarithmic scale on both axes. From *left* to *right:* English, Indonesian, Modern Greek and Russian. Same ranges along the axes for all languages.

Table 1 shows the number of analogical grids produced in each language. These numbers show that English produced the lowest number of analogical grids. Indonesian produced twice as many tables as English. Modern Greek and Russian produced five times more tables than English. Modern Greek produced a larger amount of analogical grids than Russian despite its lesser number of analogical clusters. To summarize, languages with poorer morphology tend to produce less analogical grids than languages with richer morphology, which meets intuition.

Let us recall that, by construction, on the contrary to many previous works in morphological induction [11, 5, 3, etc.], our analogical grids do not contain in any way information about word frequency, word context, nor the frequency or distribution of morphemes or the like.

3.3 Size and saturation of analogical grids

The graphs at the bottom of Figure 5 show the number of analogical grids with the same sizes in each language. Most of the analogical grids have a small size. The number of analogical grids with the same size decreases gradually as the size increases. Languages with a richer morphology produce bigger analogical

grids in average and also more analogical grids for a given size. All of this meets intuition.

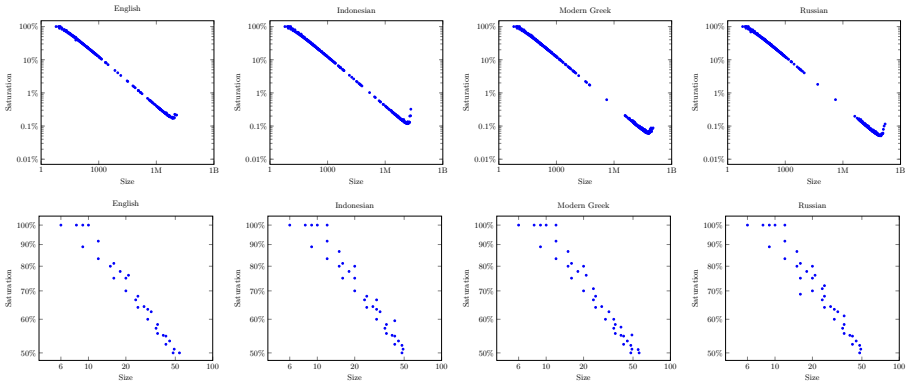

Fig. 6. Saturation of analogical grids against size in each language. From *left* to *right:* English, Indonesian, Modern Greek and Russian. Algorithmic scale on both horizontal (size) and vertical (saturation) axes. Saturation (in ordinates) in the range $[0\,\%, 100\,\%]$ *(top)* and in the range $[50\,\%, 100\,\%]$ *(bottom)*. Same ranges along the horizontal axes for all languages for the same range of saturation.

We now turn to the study of the saturation of analogical grids compared to their size. The top of Figure 6 shows saturation against size for analogical grids in each language. Analogical grids with smaller sizes tend to have higher saturation. Some tables are extremely sparse. Because of the logarithmic scale on the y-axis, the bottom half is for tables with a saturation less than 1 %.

In all cases, the plots exhibit a similar linear shape in logarithmic scale across all languages. This would correspond to Formula (5). We confirmed the similarity by the computation of the coefficients a and b for each language, as obtained by the least squares method. These coefficients are presented in Table 2. They are almost the same in all languages.

$$\log(\text{saturation}) = a \times \log(\text{size}) + b \tag{5}$$

As mentioned in Section 2.2, intuitively, analogical grids with higher saturation are more reliable to fill in because there are more word forms around the empty cells as supporting evidence. However, it may not always be the case. For instance, an analogical grid for regular English verbs extracted from any text is very hollow but empty cells can be filled in a reliable way.

4 Discussion and further experiments

Let us make a first remark on the type of the observed relation. This is not yet another instance of a Zipfian law, because, in the present case, the objects are

not ranked individually according to their frequency (number of occurrences). In a Zipfian law, the x-axis stands for the list of individual objects ranked by frequency. Recall also that our analogical grids do not encapsulate any information about the frequency of individual words whatsoever. In our graphs, two analogical grids with the same size have the same abscissa. If they also have the same saturation, they have the same ordinate and are thus plotted as the same point.

Language	Data and size	Range for saturation			
		[0%,100 %]		[50%,100 %]	
		a	b	a	b
English	Bible 100.0 %	-0.480	0.510	-0.366	0.332
	50.0 %	-0.479	0.507	-0.372	0.343
	25.0 %	-0.476	0.499	-0.368	0.336
	12.5 %	-0.474	0.491	-0.361	0.323
	Europarl (same size as Bible)	-0.481	0.516	-0.365	0.333
Indonesian	Bible 100.0 %	-0.481	0.518	-0.371	0.343
Modern Greek	”	-0.479	0.514	-0.369	0.342
Russian	”	-0.482	0.520	-0.370	0.342

Table 2. Linear coefficients for each language; and for different sizes and different genres in English.

The interesting fact that comes into light is not so much the fact that the relation between size and saturation of analogical grids be a log–log relation, but the fact that it exhibits very similar slopes in all four languages. A reasonable explanation is that these coefficients are independent of the language because they characterize the corpus used. The corpus is defined by its size and its genre.

We first inquired whether the coefficients depend on the size of the corpus used. We performed the same experiment in English and let the size of the corpus vary: a half, a quarter, an eighth of the original size. The computation of the coefficients led to very similar results as shown in Table 2.

We then inquired the influence of the genre and performed the same experiment with the same size of text in English again. We chose the Europarl corpus for this experiment. Again, the computation of the linear coefficients led to very similar results, as shown in Table 2.

Further experiments with more parameters varying are required to confirm that the coefficients of the relationship between saturation and the size are always very similar. However, for the time being, we observe that the parameters are relatively close at least for these four languages whith different richness in morphology.

5 Conclusion

We studied analogical grids in different languages with different morphological richness. These analogical grids were automatically built from actual texts, using a technique which has been presented in previous work. Without surprise, languages known to be richer in morphology produce bigger and more analogical grids than languages less rich in morphology. Empty cells in such analogical grids are interesting because they could be filled by words that should then be tested against the actual language.

We studied the relation between size and saturation in analogical grids. Experimental results clearly showed that the logarithm of the saturation of an analogical grids linearly depends on the logarithm of its size. This is not so surprising. More interestingly, the computation of the coefficients characterizing this log-log linear relation led to the result that, across all the four languages used, and even when having size and genre varying in one language, these coefficients are almost always the same: the relation between the saturation and the size of an analogical grid would be almost independent of the size, the genre and the language of a text.

References

1. Beesley, K.R.: Consonant spreading in Arabic stems. In: Proceedings of COLING-ACL'98. vol. I, pp. 117–123. Montréal (Aug 1998), http://www.aclweb.org/anthology/P98-1018
2. Chan, E.: Structures and distributions in morphology learning. Ph.D. thesis, University of Pennsylvania. (2008), http://nlp.cs.swarthmore.edu/~richardw/papers/chan2008-structures.pdf
3. Dryer, M., Eisner, J.: Discovering morphological paradigms from plain text using a dirichlet process mixture model. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'2011). pp. 616–627. Association for Computational Linguistics, Edinburgh, Scotland, UK (2011), https://www.cs.jhu.edu/~jason/papers/dreyer+eisner.emnlp11.pdf
4. Fam, R., Lepage, Y.: Morphological predictability of unseen words using computational analogy. In: Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-CA-16). pp. 51–60. Atlanta, Georgia (2016)
5. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics 27, 153–198 (2001)
6. Hathout, N.: Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In: Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing. pp. 1–8. Coling 2008 Organizing Committee, Manchester, UK (August 2008), http://www.aclweb.org/anthology/W08-2001
7. Langlais, P., Yvon, F.: Scaling up analogical learning. In: Coling 2008: Companion volume: Posters. pp. 51–54. Coling 2008 Organizing Committee, Manchester, UK (August 2008), http://www.aclweb.org/anthology/C08-2013

8. Lepage, Y.: Analogies between binary images: Application to Chinese characters. In: Prade, H., Richard, G. (eds.) Computational Approaches to Analogical Reasoning: Current Trends, pp. 25–57. Springer, Berlin, Heidelberg (2014), http://dx.doi.org/10.1007/978-3-642-54516-0_2
9. Neuvel, S., Fulop, S.A.: Unsupervised learning of morphology without morphemes. In: Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning. pp. 31–40. Association for Computational Linguistics (July 2002), http://www.aclweb.org/anthology/W02-0604
10. Resnik, P., Olsen, M.B., Diab, M.: The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. Computers and the Humanities 33(1), 129–153 (1999), http://dx.doi.org/10.1023/A:1001798929185
11. Schone, P., Jurafsky, D.: Knowledge-free induction of morphology using latent semantic analysis. In: Proceedings of CoNLL-2000 and LLL-2000. pp. 67–72. Lisbon, Portugal (2000), http://web.stanford.edu/~jurafsky/W00-0712.pdf
12. Singh, R., Ford, A.: In praise of Sakatayana: some remarks on whole word morphology. In: Singh, R. (ed.) The Yearbook of South Asian Languages and Linguistics-200. Sage, Thousand Oaks (2000)
13. Stroppa, N., Yvon, F.: An analogical learner for morphological analysis. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). pp. 120–127. Association for Computational Linguistics, Ann Arbor, Michigan (June 2005), http://www.aclweb.org/anthology/W/W05/W05-0616
14. Wintner, S.: Natural Language Processing of Semitic Languages, chap. Morphological Processing of Semitic Languages, pp. 43–66. Springer, Berlin, Heidelberg (2014), http://dx.doi.org/10.1007/978-3-642-45358-8_2