information

MDPI

Article

A Study of Analogical Density in Various Corpora at Various Granularity

Rashel Fam *,† and **Yves Lepage** †

Graduate School of IPS, Waseda University, Kitakyushu-shi, Fukuoka-ken 808-0135, Japan;
yves.lepage@waseda.jp
* Correspondence: fam.rashel@fuji.waseda.jp; Tel.: +81-80-2776-8224
† 2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka-ken 808-0135, Japan.

Abstract: In this paper, we inspect the theoretical problem of counting the number of analogies between sentences contained in a text. Based on this, we measure the analogical density of the text. We focus on analogy at the sentence level, based on the level of form rather than on the level of semantics. Experiments are carried on two different corpora in six European languages known to have various levels of morphological richness. Corpora are tokenised using several tokenisation schemes: character, sub-word and word. For the sub-word tokenisation scheme, we employ two popular sub-word models: unigram language model and byte-pair-encoding. The results show that the corpus with a higher Type-Token Ratio tends to have higher analogical density. We also observe that masking the tokens based on their frequency helps to increase the analogical density. As for the tokenisation scheme, the results show that analogical density decreases from the character to word. However, this is not true when tokens are masked based on their frequencies. We find that tokenising the sentences using sub-word models and masking the least frequent tokens increase analogical density.

Keywords: proportional analogy; language productivity; automatic acquisition

check for updates

Citation: Fam, R.; Lepage, Y. A Study of Analogical Density in Various Corpora at Various Granularity. *Information* **2021**, *12*, 314. https://doi.org/10.3390/info12080314

Academic Editor: Kostas Stefanidis

Received: 21 June 2021
Accepted: 23 July 2021
Published: 5 August 2021

1. Introduction

Analogy is a relationship between four objects that states the following: A is to B as C is to D. When the objects are pieces of a text, analogies can be of different sorts:

- World-knowledge or pragmatic sort, as in
  Indonesia : *Jakarta* :: *Brazil* : *Brasilia*
  (state/capital);
- Semantic sort, as in
  glove : *hand* :: *envelope* : *letter*
  (container/content);
- Grammatical sort, as in
  child : *children* :: *man* : *men*
  (singular/plural); and
- Formal sort, or level of form, as in
  he : *her* :: *dance* : *dancer* (suffixing with r).

In this work, we focus on analogy on the level of form. Analogies on the level of form have been used in morphology to extract analogical grids, i.e., types of paradigm tables that contain only regular forms [1–5]. The empty cells of such grids can be filled in with new words, known as the lexical productivity of language [6–8]. The reliability of newly generated words can be assessed by various methods [9,10].

1.1. Motivation and Justification

Analogies on the level of form have also been used between sentences in an example-based machine translation (EBMT) system [11]. The reported system used a very particular corpus, the Basic Traveler's Expression Corpus (BTEC) presented in [12], where sentences are very short (average length of eight words for English) and where similar sentences are very frequent. For these reasons, the chances of finding analogies in that corpus were high [13]. Such conditions seriously limited the application of the method and prevented its application to more standard corpora such as the Europarl corpus [14], where sentences have an average length of 30 words (for English) and where similar sentences are not frequent.

The higher number of analogies in the corpus intuitively increases the chances of translation with the EBMT system. In the EBMT system, translations are made using the target sentence and other sentences contained in the knowledge database. In this way, we explain how the translation was made without any meta-language, such as parts of speech or parse trees. The translation was explained by the language itself through examples (facts) contained in the data. Thus, we want to have as many sentences as possible be covered by analogies. Furthermore, sentences that are not contained in the training data or knowledge database, called external sentences, are expected to be covered if they have the same style or domain.

1.2. Contributions

The present paper examines in more details the number of formal analogies that can be found between sentences in various corpora in various languages. For that purpose, the notion of analogical density is introduced. It allows us to inspect what granularity or what masking techniques for sentences may lead to higher analogical densities.

The contributions of this work are thus summarised as follows:

- We introduce a precise notion of analogical density and measure the analogical density of various corpora;
- We characterise texts that are more likely to have a higher analogical density;
- We investigate the effect of using different tokenisation schemes and the effect of masking tokens by their frequency on the analogical density of various corpora;
- We investigate the impact of the average length of sentences on their analogical density of corpora; and
- Based on previously mentioned results, we propose general rules to increase the analogical density of a given corpus.

1.3. Organisation of the Paper

The paper is organised as follows: Sections 2 and 3 introduce the basic notions of formal analogy and analogical density. Section 4 presents the data and several statistics. Section 5 introduce several methods used to tokenise the text. Section 6 explains how we mask the tokens based on their frequency to further boost the analogical density. Section 7 presents the experimental protocol and results. Sections 8 and 9 give further discussions on the results and conclusion of this work.

2. Number of Analogies in a Text and Analogical Density

We address the theoretical problem of counting the total number of analogies contained in a given text. In the following section, we introduce two main metrics used in this work.

2.1. Analogical Density

The analogical density (D_{nlg}) of a corpus is defined as the ratio of the total number of analogies contained in the corpus (N_{nlg}) against the total number of permutations of four objects that can be constructed by the number of sentences (N_s).

$$D_{nlg} = \frac{N_{nlg}}{\frac{1}{8} \times N_s^4} = 8 \times \frac{N_{nlg}}{N_s^4} \tag{1}$$

The factor 1/8 in the denominator comes from the fact that there exist eight equivalent forms of one same analogy due to the two main properties of an analogy:

- symmetry of conformity: $A : B :: C : D \Leftrightarrow C : D :: A : B$;
- exchange of the means: $A : B :: C : D \Leftrightarrow A : C :: B : D$.

This is illustrated in Figure 1. Please see below, in Section 3.1, for further details.

$A : B :: C : D$	
$A : C :: B : D$	**exchange of the means**
$C : D :: A : B$	**symmetry of conformity**
$B : D :: A : C$	
$C : A :: D : B$	
$B : A :: D : C$	(inversion of ratios)
$D : B :: C : A$	(exchange of the extremes)
$D : C :: B : A$	(inversion of reading)

Figure 1. The eight equivalent forms of a same analogy $A : B :: C : D$.

2.2. Proportion of Sentences Appearing in Analogy

We count the number of sentences appearing in at least one analogy (N_{s_nlg}) and take the ratio with the total number of sentences in the corpus (N_s) to obtain the proportion of sentences appearing in at least one analogy (P).

$$P = \frac{N_{s\_nlg}}{N_s} \tag{2}$$

2.3. Meaning of the Measure and Gauging

A value of 1 for density means that the set of strings is reduced to a singleton. Usually, values of density are very low and are better expressed in centi (c), mili (m), micro (μ), nano (n), or even pico (p) and femto (f). See SI units in Table 1. This is intuitive when we examine Formula (1). The denominator grows exponentially with the number of sentences (permutation of 4). This implies that the analogical density is very low as the number of analogies will be a lot less because one has to satisfy the constraint (commutation).

Table 1. Small units in the International System if Units (SI).

Prefix	Symbol	Factor
centi	c	10^{-2}
mili	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}
zepto	z	10^{-21}
yocto	y	10^{-24}

2.4. Restrictions

We restrict ourselves to the case of counting the analogies between a string of a given size. It is the third kind of analogy presented in Section 1. In this work, we focus on the level of form, not on the level of semantics. The object of the analogy that we work with is not words but sentences. We observe the commutation of different kinds of units: character, sub-word and word. Please refer to Section 5 for further details about how we tokenise the data. As for the definition of analogy, in this paper, we adopt the definition of formal analogies between strings of symbols found in [15–17].

3. Analogy

Analogy is a relationship between four objects: A, B, C and D where A is to B as C is to D. It is noted as $A : B :: C : D$. As our work relate with strings, A, B, C and D are all strings (sequence of characters). This notation means that the ratio between A and B is similar to the ratio between C and D. In other words, analogy is a conformity of ratios between the four strings. Figure 2 gives examples of analogies between sentences.

$$\begin{array}{ll} \textit{I like coffee} : \textit{I like tea} :: \textit{I like hot coffee} : \textit{I like hot tea} & \\ \Leftrightarrow & \\ \textit{I like hot coffee} : \textit{I like hot tea} :: \textit{I like coffee} : \textit{I like tea} & (\textbf{symmetry of conformity}) \\ \Leftrightarrow & \\ \textit{I like coffee} : \textit{I like hot coffee} :: \textit{I like tea} : \textit{I like hot tea} & (\textbf{exchange of the means}) \end{array}$$

Figure 2. Examples of analogies in sentences with equivalent analogies derived from the properties of analogies mentioned in Section 3.1.

3.1. Properties of Analogy

There are three general properties of analogies.

- Reflexivity of conformity
  $A : B :: A : B$ is always a valid analogy for any A and B.
  $$\textit{slow} : \textit{slower} :: \textit{slow} : \textit{slower}$$
- Symmetry of conformity
  Let $A : B :: C : D$ be a valid analogy; then, equivalently, $C : D :: A : B$ is also valid.
  $$\textit{slow} : \textit{slower} :: \textit{high} : \textit{higher} \quad \Leftrightarrow \quad \textit{high} : \textit{higher} :: \textit{slow} : \textit{slower}$$
- Exchange of the means
  Let $A : B :: C : D$ be a valid analogy; then, equivalently, $A : C :: B : D$ is also valid.
  $$\textit{slow} : \textit{slower} :: \textit{high} : \textit{higher} \quad \Leftrightarrow \quad \textit{slow} : \textit{high} :: \textit{slower} : \textit{higher}$$

Based on the last two properties, we can understand that there are eight equivalent forms for one valid analogy. In this work, we also excluded considerations of the property of reflexivity of conformity. Figure 1 shows that the eight equivalent forms for $A : B :: C : D$ is a valid analogy. As this paper focuses on analogies between sentences, Figure 2 shows analogies in sentences.

3.2. Ratio between Strings

To define the ratio between strings, we need to first define how to represent strings. We consider representing each sentence using the vector shown in Formula (3). The features are the number of occurrences of each token or character in the sentence. Formula (3) illustrates the Parikh vector of a string that uses the number of occurrences of each character in a string. If we tokenise the sentence by words, then the feature is the number of occurrences of each words. We use the notation $|S|_c$, which stands for the number of occurrences of a character or token c in string S. The number of dimensions of the vector is the size of the alphabet or the vocabulary depending on the tokenisation scheme.

$$A \triangleq \begin{pmatrix} |A|_a \\ |A|_b \\ \vdots \\ |A|_s \\ \vdots \\ |A|_z \end{pmatrix} \qquad example = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3}$$

From this, we define the ratio between strings as the difference between the string representation and its edit distance. Formula (4) defines the ratio between two strings

A and B. Notice the difference in the features of character s for the ratio between the word 'example' and 'examples'. The differences in the number of occurrences for all characters or tokens come from the characterisation of proportional analogy in [16] or [17]. The last dimension, written as $d(A, B)$, is the LCS edit distance between the two strings. This indirectly gives the number of common characters appearing in the same order in A and B. 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.

The above definition of ratios captures prefixing and suffixing. Although we do not show it here, this definition also captures parallel infixing or interdigitation, a well-known phenomenon in the morphology of semitic languages [18,19]. However, partial reduplication (e.g., consonant spreading) or total reduplication [20] (e.g., marked plural in Indonesian) are not captured by this definition.

$$A : B \triangleq \begin{pmatrix} |A|_a - |B|_a \\ |A|_b - |B|_b \\ \vdots \\ |A|_s - |B|_s \\ \vdots \\ |A|_z - |B|_z \\ d(A,B) \end{pmatrix} \qquad example : examples = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ -1 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{4}$$

3.3. Conformity of Ratios between Strings

The conformity between ratios of strings is defined as the equivalent between the two vectors of ratios. See Formula (5).

$$A : B :: C : D \quad \overset{\Delta}{\Longleftrightarrow} \quad \begin{cases} A : B & = & C : D \\ A : C & = & B : D \end{cases} \tag{5}$$

This definition confirms the properties of analogy mentioned in Section 3.1. In this way, we ensure that the use of vector representation of strings satisfies the properties of analogy. These properties are also carried to the definition of analogical cluster.

3.4. Analogical Cluster: Cluster of Similar Ratios

Pairs of strings representing the same ratio can be grouped as an analogical cluster. Please refer to Formula (6). Notice that the order of string pairs has no importance.

$$\begin{matrix} A_1 : B_1 \\ A_2 : B_2 \\ \vdots \\ A_n : B_n \end{matrix} \quad \overset{\Delta}{\Longleftrightarrow} \quad \begin{array}{l} \forall (i,j) \in \{1, \ldots, n\}^2, \\ \qquad A_i : B_i :: A_j : B_j \end{array} \tag{6}$$

We compute all ratios between strings and then group string pairs that represents the same ratio. Ideally, we have to compute the all of the ratios directly. However, it is a very time consuming and exhaustive task. Here, we adopted the two-step approach proposed in [21] for analogies between binary images.

The idea is to first represent the set of all sentences as a tree. Each level in the tree stands for a token contained in the vocabulary. Then, the sentences are hierarchically grouped based on the number of occurrences of the tokens. The tree representation is explored in a top-down manner against the tree itself. The purpose is to group string pairs by equal difference in the number of occurrences of tokens.

Finally, we verify the distance of each string pairs in two ways:

- horizontally: between A_i and B_i;

- vertically: between $A_i : B_i$ and $A_j : B_j$.

Due to this, we may split the group of string pairs into smaller groups if we find a difference in distance.

4. Survey on the Data

The experiments were carried on six European languages, ranked by the order of morphological richness: English (en), French (fr), German (de), Czech (cs), Polish (pl) and Finnish (fi). We considered surveying four different corpora that cover the previously mentioned languages. Most of these corpora are mainly used in machine translation tasks. Table 2 shows the language availability for each corpus. Below are the corpora ranked in the decreasing order of expected analogical density:

- Tatoeba (available at: tatoeba.org accessed on 20 September 2020) is a collection of sentences that are translations provided through collaborative works online (crowdsourcing). It covers hundreds of languages. However, the amount of data between languages are not balanced because it also depends on the number of members who are native speakers of that language. Sentences contained in Tatoeba corpus are usually short. These sentences are mostly about daily life conversations. Table 3 shows the statistics of Tatoeba corpus used in the experiments.
- Multi30K (available at: github.com/multi30k/dataset accessed on 20 September 2020) [22–24] is a collection of image descriptions (captions) provided in several languages. This dataset is mainly used for multilingual image description and multimodal machine translation tasks. It is an extension of Flickr30K [25], and more data are added from time to time, for example, the COCO dataset (available at: cocodataset.org accessed on 20 September 2020). Table 4 shows the statistics of Multi30K corpus.
- CommonCrawl (available at: commoncrawl.org accessed on 20 September 2020) is a crawled web archive and dataset. Due to its nature as web archives, this corpus covers a lot of topics. In this paper, we used the version that is provided as training data for the Shared Task: Machine Translation of WMT-2015 (available at: statmt.org/wmt15/translation-task.html accessed on September 2020). Table 5 shows the statistics on the CommonCrawl corpus.
- Europarl (available at: statmt.org/europarl/ accessed on September 2020) [14] is a corpus that contains transcriptions of the European Parliament in 11 European languages. It was first introduced for Statistical Machine Translation and is still used as the basic corpus for machine translation tasks. In this paper, we use version 7. Table 6 shows the statistics on Europarl corpus.

Table 2. Languages provided by each corpus used in the experiment.

	en	**fr**	**de**	**cs**	**pl**	**fi**
Tatoeba	✓	✓	✓		✓	✓
Multi30K	✓	✓	✓	✓		
CommonCrawl	✓	✓	✓	✓		
Europarl	✓	✓	✓		✓	✓

Table 3. Statistics on Tatoeba corpus.

-	**en**	**fr**	**de**	**cs**	**pl**	**fi**
# of lines	7964					
# of tokens	51,279	54,430	50,375	-	41,892	39,907
# of types	4152	5740	5639	-	7796	8634
Avg. tokens per line	6.44 ± 2.80	6.83 ± 3.20	6.33 ± 2.85	-	5.26 ± 2.44	5.01 ± 2.10
Avg. token length	3.34 ± 2.14	3.69 ± 2.52	4.04 ± 2.59	-	4.26 ± 2.89	4.75 ± 3.15
Avg. type length	6.32 ± 2.27	7.12 ± 2.49	7.55 ± 3.01	-	7.28 ± 2.44	8.09 ± 2.86
Type-Token-Ratio	0.08	0.11	0.11	-	0.19	0.22
Hapax size (%)	48.80	56.17	55.68	-	62.11	66.50

Table 4. Same as Table 3 but on the Multi30K corpus.

-	**en**	**fr**	**de**	**cs**	**pl**	**fi**
# of lines	30,014					
# of tokens	392,978	471,352	374,490	308,367	-	-
# of types	10,373	11,376	19,112	22,787	-	-
Avg. tokens per line	13.09 ± 4.10	15.70 ± 5.91	12.48 ± 4.23	10.27 ± 3.60	-	-
Avg. token length	3.85 ± 2.40	3.93 ± 2.47	4.86 ± 2.97	4.34 ± 2.71	-	-
Avg. type length	6.92 ± 2.41	7.41 ± 2.42	9.91 ± 3.91	7.52 ± 2.40	-	-
Type-Token-Ratio	0.03	0.02	0.05	0.07		
Hapax size (%)	41.94	42.15	58.05	53.50	-	-

Table 5. Same as Table 3 but on the CommonCrawl corpus.

-	**en**	**fr**	**de**	**cs**	**pl**	**fi**
# of lines	27,379					
# of tokens	582,530	647,747	539,655	548,214	-	-
# of types	27,880	33,592	43,073	47,788	-	-
Avg. tokens per line	21.28 ± 13.08	23.66 ± 16.71	19.71 ± 15.15	20.02 ± 15.43	-	-
Avg. token length	4.40 ± 2.84	4.54 ± 3.20	5.29 ±4.08	4.70 ± 3.14	-	-
Avg. type length	7.30 ± 3.06	7.53 ± 3.03	9.34 ± 4.55	7.53 ± 2.81	-	-
Type-Token-Ratio	0.05	0.05	0.08	0.09	-	-
Hapax size (%)	49.15	49.91	56.93	53.09	-	-

Table 6. Same as Table 3 but on the Europarl corpus.

-	**en**	**fr**	**de**	**cs**	**pl**	**fi**
# of lines	186,303					
# of tokens	5,596,191	6,195,568	5,401,483	-	4,805,319	3,963,855
# of types	39,042	52,662	108,526	-	108,777	188,718
Avg. tokens per line	30.04 ± 15.51	33.26 ± 17.42	28.99 ± 15.12	-	25.79 ± 13.60	21.28 ± 10.97
Avg. token length	4.54 ± 2.98	4.66 ± 3.25	5.55 ± 4.04	-	5.70 ± 3.85	6.90 ± 4.54
Avg. type length	8.37 ± 3.37	8.67 ± 3.11	12.66 ± 5.32	-	9.61 ± 3.24	12.83 ± 4.78
Type-Token-Ratio	0.01	0.01	0.02	-	0.02	0.05
Hapax size (%)	36.08	36.12	50.37	-	39.81	52.62

Europarl emerges as the corpus with the highest number of lines. It also has the highest average number of tokens per line. In contrast, Tatoeba has the smallest average number of tokens per line, as expected. As an overview, Multi30K has two times the number of tokens in a sentence in comparison with Tatoeba, and CommonCrawl has three times while Europarl has around four times. Our hypothesis is that tokens in shorter sentences have more chances to commute. Thus, it has more analogies.

These four corpora can be characterised into two groups based on the diversity of the sentence context. Multi30K and CommonCrawl are corpora with diverse contexts. In comparison with that, sentences contained in Tatoeba and Europarl are less diverse. Tatoeba is mostly about daily life conversation, while Europarl is a discussion on parliament. We expect that corpora with less diversity in their context share words between sentences more often. Thus, it has more analogies and a higher analogical density.

Let us now compare the statistics between languages. English has the lowest number of types. Finnish, Polish and Czech always have the highest number of types, around two times higher than English across the corpora. It is even more than four times higher for Europarl. We can observe that languages with poor morphology have fewer of types and hapaxes. On the contrary, languages with high morphological richness have less number of tokens due to a richer vocabulary. These languages also tend to have longer words (in characters). One can easily understand that with richer morphological features, we have larger vocabulary. The consequence of this is that the words are longer. We also observe that a higher number of types means less words to repeat (higher Type–Token Ratio). Thus, the number of tokens is lower.

However, we also see that there are some interesting exceptions, in this case, French and German. French has a higher number of tokens than English despite having higher vocabulary size. The greater variety in the number of functional words (propositions, articles, etc.) in French is probably one of the explanations for this phenomenon. As for German, it has a pretty high average length of type in comparison with other languages. This is perhaps caused by words in German being originally longer. German is known to glue several words into a compound word. Table 7 provides example of sentences contained in the corpora.

Table 7. Example sentences (lowercased and tokenised) randomly chosen from the corpora used in the experiment. Sentences contained in the same corpus are translations of each other in other languages.

Corpus	Lang.	Example Sentences
Tatoeba	en	*the store is closing at 7.*
	fr	*le magasin ferme à 7 heures.*
	de	*der laden schließt um sieben.*
	pl	*sklep jest zamknięty od 19.*
	fi	*kauppa menee kiinni kello seitsemän.*
Multi30K	en	*a boy in white plays baseball.*
	fr	*un garçon en blanc joue au baseball.*
	de	*ein weiß gekleideter junge spielt baseball.*
	cs	*chlapec v bílém hraje baseball.*
CommonCrawl	en	*remember—this is a brief selection of photographs—much more you can see the individual entries on the blog.*
	fr	*n' oubliez pas—il s' agit d' une brève sélection de photographies—bien plus que vous pouvez voir les entrées individuelles sur le blog.*
	de	*denken sie daran—das ist eine kleine auswahl von fotografien—viel mehr können sie die einzelnen einträge auf dem blog zu sehen.*
	cs	*pamatujte si—to je stručný výběr fotografií—mnohem více můžete vidět jednotlivé položky na blogu.*
Europarl	en	*(de) madam president, in terms of european integration, it is without doubt a good thing that one of the new eu countries, in this case the czech republic, held the council presidency.*
	fr	*(de) madame la présidente, en termes d' intégration européenne, il est certes une bonne chose que la présidence du conseil revienne à l' un des nouveaux états membres de l' ue, en l' espèce, la république tchèque.*
	de	*(de) frau präsidentin ! im sinne der europäischen integration war es zweifellos begrüßenswert, dass mit tschechien eines der neuen eu-länder die ratspräsidentschaft innehatte.*
	pl	*(de) pani przewodnicząca ! jeżeli chodzi o integrację europejską, bez wątpienia dobrą rzeczą jest to, że jeden z nowych krajów, a mianowicie republika czeska, sprawował prezydencję rady.*
	fi	*(de) arvoisa puhemies, euroopan yhdentymisen kannalta on epäilemättä hyvä asia, että yksi eu:n uusista jäsenvaltioista, tässä tapauksessa tšekin tasavalta, toimi neuvoston puheenjohtajana.*

Aligning Sentences across Languages

For Europarl and Multi30K, there exist parallel corpora. However, some corpora are not aligned, in this case, Tatoeba and CommonCrawl. For these corpora, we need to align the sentences contained in the corpus. Having parallel corpora allows us to make a comparison between languages.

For each corpus, we used English as the pivot language to align the sentences across the other languages. We added an English sentence to the collection of aligned sentences if the sentence has translations in the other languages. If there were several translation references are available in another language, one sentence was randomly picked to represent that particular language. Thus, for each English sentence, there is only one corresponding sentence in every language at the end of the alignment process.

5. Tokenisation

As a reminder, the notion of analogy considered in this paper is that of analogy of commutation between strings. Thus, we considered several approaches to tokenise the corpora: character, sub-word and word. All of the corpora were first preprocessed using the preprocessing script MOSES (available at: github.com/moses-smt/mosesdecoder accessed on 20 September 2020). Table 8 gives examples of different tokenisation schemes.

Table 8. Examples of different tokenisations on the the same sentence taken from the Tatoeba corpus. The sentence is tokenised using different tokenisation schemes: character, sub-word and word. For sub-words, we used two popular sub-word models: unigram language model (unigram) and byte-pair encoding (BPE). The delimiter used to separate tokens is the space. Underscores denote spaces in the original sentence. The vocabulary size used here for unigram and BPE is 1000 (1 k).

Original	***The Store Is Closing at 7.***
character	*t h e _ s t o r e _ i s _ c l o s i n g _ a t _ 7 _ .*
unigram	*_the _stor e _is _c l o s ing _at _ 7 _.*
BPE	*_the _st ore _is _cl os ing _at _ 7 _.*
word	*the store is closing at 7 .*

5.1. Character

We consider the character to be the most basic unit used. White spaces (spaces, tabulations and newlines) are annotated as underscore for us to know where the word boundaries are. Our hypothesis is that the commutation of characters between sentences is relatively easier to observe in comparison with longer sequences of characters (sub-words or words). Due to this, we expect a higher number of analogies from the corpora, with the character as the tokenisation unit.

5.2. Sub-Word

We considered two popular sub-word models to tokenise the corpora: unigram language model (unigram) [26] and byte-pair encoding (BPE) [27]. For both BPE and unigram, we used the Python module implementation provided by SentencePiece (available at: github.com/google/sentencepiece accessed on 20 September 2020). Varying the vocabulary size by 250, 500, 750, 1 k and 2 k is used to train the model for both techniques.

5.2.1. Token Length

Figure 3 shows the average length of the token and type of Tatoeba corpus for English after being tokenised using both sub-word algorithms, BPE and unigram, with different vocabulary sizes as its parameter. When the vocabulary size rises, both the BPE and unigram tokenise the corpus into a longer sequence of characters, resulting in longer tokens and types. We can also observe that, most of the time, as the vocabulary size goes up, the number of tokens decreases while the number of hapaxes increases. This is consistent with the observation we made in the previous section.

Let us also compare the difference between BPE and unigram. Unigram tends to tokenise the text into a longer sequence of characters in comparison with BPE.

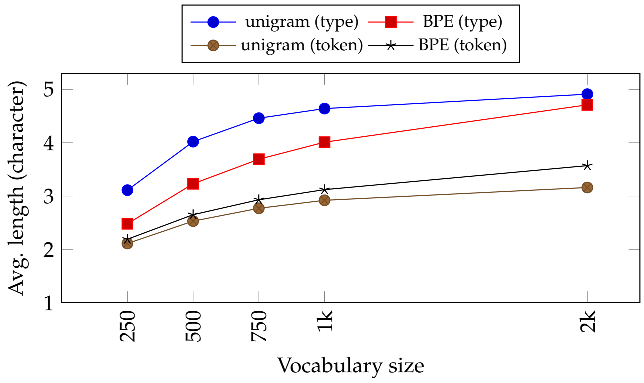


Figure 3. Average token and type lengths (in character) on the English part of the Tatoeba corpus after tokenisation using BPE and unigram with different sizes of vocabulary. We do not provide the figures with vocabulary sizes from 4 k onwards because only BPE is able to produce tokenisation with the mentioned parameter.

5.2.2. Sampling

The use of the regularisation method (sampling) is known to improve the performance and robustness of Neural Machine Translation (NMT). It is introduced in both sub-word algorithms, called sub-word regularisation [26] and BPE dropout [28]. The idea is to virtually augment the data with on-the-fly sampling.

Our preliminary experimental results show that the use of sampling when performing tokenisation, with both BPE and unigram, decreases the number of analogies extracted from a corpus. Our intuition is that this is due to the nature of randomness that is introduced when tokenising a corpus. As there is no consistent tokenisation for the same words, we hardly find the same commutation even between sentences that are very similar. Based on these results, we decided not to use sampling for further experiments.

5.3. Word

For word tokenisation, we simply used white spaces as the delimiter. This is the standard tokenisation for most natural language processing tasks.

6. Masking

We blurred the tokens by masking either the most frequent types or least frequent types. According to Zipf's law, we determined the place where the balanced is obtained. The power law states the following:

$$\mathrm{f}(w) \times \mathrm{r}(w) = C \tag{7}$$

Let us call N the total number of words. We determined the rank ρ:

$$N \times \sum_{1}^{\rho} \mathrm{f}(w) = (N - \rho) \times \sum_{1}^{N} \mathrm{f}(w) \tag{8}$$

Pareto's famous claim was that 20% or the richest own 80% of the riches.

By relying on Zipf's law, the words in a corpus are divided into two categories. This is a similar trick proposed by [29] for an approximation of EM algorithm.

In this paper, we considered using the following scenario:

- least frequent: tokens which belong to the N least frequent types (caution: tokens are repeated. Types are counted only once) are masked with one same label, while all the other types are kept as it is. In this paper, we ranked the types according to their frequency in the corpus. After that, we masked all tokens in the corpus that belong to the least frequent types for which the accumulated frequency is half of the total number of tokens in the corpus. All other tokens are kept. If several types in the same rank (frequency) exist, then we just keep randomly picking one of them until the accumulated frequency is half of the total number of tokens.

- most frequent: same as above but with the token with N most frequent types instead (opposite of the least frequent).

We expect to see an increase in analogical density by masking the tokens, especially when the least frequent tokens are masked. Under this condition, the sentences are contained with mostly functional words with masked slots. These functional words show the structure of the sentence. Table 9 presents examples of the masking performed on both word and sub-word tokenisation schemes.

Table 9. An example of masking a sentence contained in the Tatoeba corpus for the tokenisation scheme on word (top) and BPE sub-word (bottom). In this example, we use the same label '**...**' to replace the most or least frequent types. The example sentence is the same sentence shown in Table 8.

Tokenisation	**Masking**	**Example Sentence**										
word	-	*the*	*store*	*is*	*closing*	*at*	*7*	.				
	least	*the*	...	*is*	...	...	...	.				
	most	...	*store*	...	*closing*	*at*	*7*	...				
BPE	-	*_the*	*_st*	*ore*	*_is*	*_cl*	*os*	*ing*	*_at*	_	*7*	_.
	least	*_the*	...	...	*_is*	...	...	*ing*	...	...	...	_.
	most	...	*_st*	*ore*	...	*_cl*	*os*	...	*_at*	_	*7*	...

7. Results and Analysis

7.1. Effect of Tokenisation on Analogical Density

Each of the corpora is tokenised using four different tokenisation schemes: character, BPE. unigram and word. On top of that, we performed masking with both methods: the least frequent and most frequent. Ablation experiments are carried on all corpora in six languages depending on the language availability of the corpus.

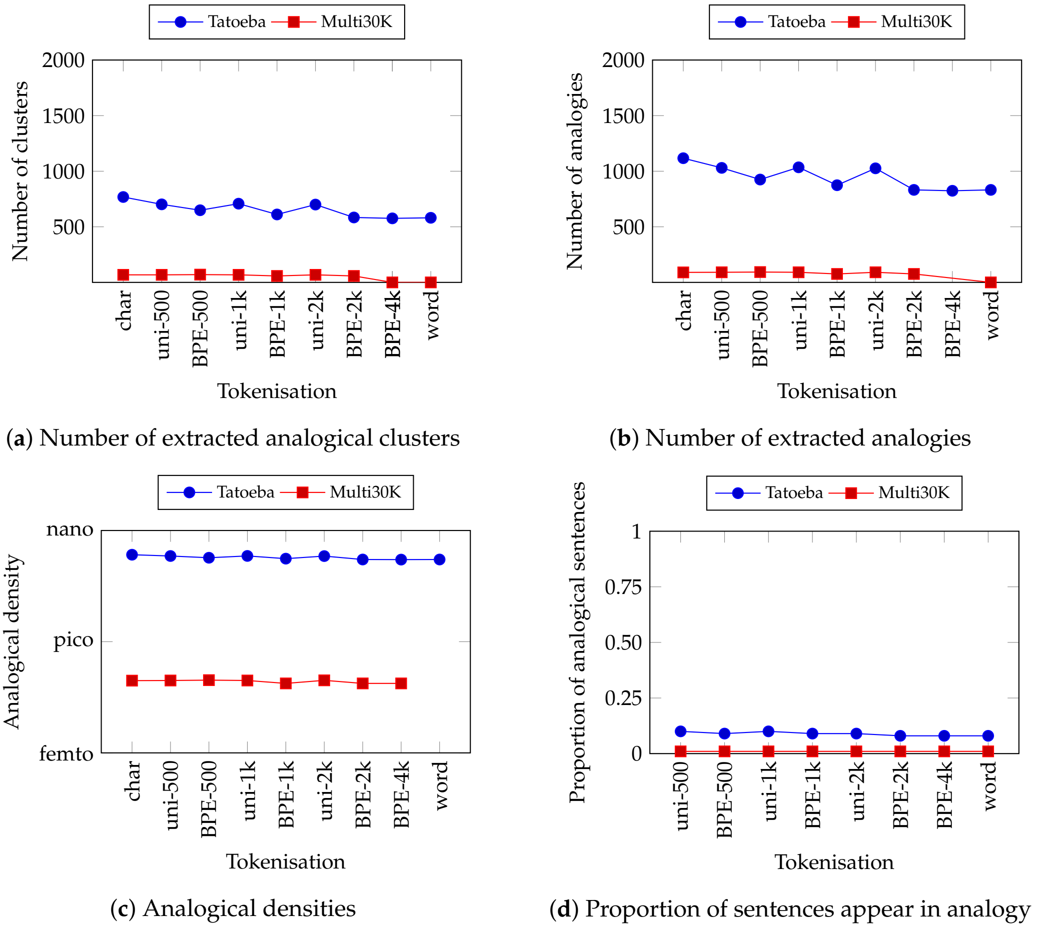
In this paper, we decided to carry the experiment on both Tatoeba and Multi30K as these corpora have a different range on both the formal level and the semantic level. On the formal level, sentences in Tatoeba are short and similar to one another. Multi30K contains more diverse and longer sentences. On the level of semantics, as mentioned in Section 4, Tatoeba contains sentences that focus on the theme of daily conversation. Multi30K, which contains image captions, has a wider range of topics.

Figure 4a (top-left) shows the number of analogical clusters extracted from the corpora with various tokenisations in English. Tatoeba has the highest number of clusters. This meets our hypothesis. It is also reflected in the number of analogies shown in Figure 4b (top-right). Tatoeba has about 10 times more analogies than Multi30K.

Figure 4c (bottom-left) shows the results on the analogical density of the corpora with various tokenisations. We can immediately observe that the Tatoeba corpus steadily has the highest analogical density in comparison with the other corpora. The difference is also pretty far. For example, the gap is around 10^3 between Tatoeba and Multi30K, even more than 10^5 for Europarl. This shows that the Tatoeba corpus is really more dense than the other corpora despite having the smallest number of sentences. Remember, we have a different number of sentences between corpora.

Although it is not visible from the graph, we observed that the density slightly decreases from tokenisation in character towards words. For subword tokenisation, we found that unigram consistently has higher analogical density than BPE on the same vocabulary size. This is probably caused by unigram having a shorter token length, which allows for a higher degree of freedom in commutation between tokens.

Similar trends can also be observed in the proportion of analogical sentences. This is shown in FIgure 4d (bottom-right). Tatoeba is ten times higher than Multi30K which proves our hypothesis that a corpus containing similar sentences has a higher proportion of analogical sentences. As for the tokenisation scheme, we also found that the proportion decreases toward word tokenisation.



(**a**) Number of extracted analogical clusters

(**b**) Number of extracted analogies

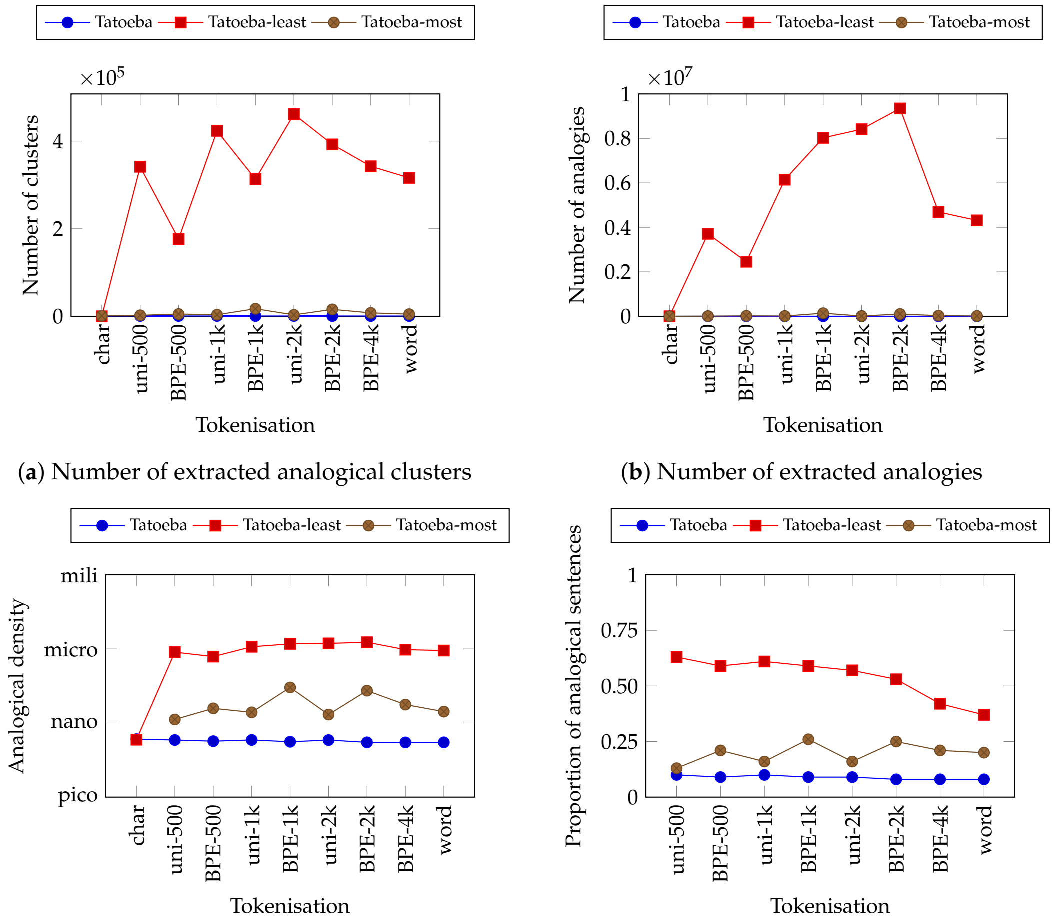
(**c**) Analogical densities

(**d**) Proportion of sentences appear in analogy

Figure 4. Number of clusters (**a**) and analogies (**b**) extracted from the corpora in English. Below that, Analogical density (**c**) and the proportion of sentences appear in analogy (**d**) for the corpora in English. Please take note of the logarithmic scale on the ordinate for analogical density (micro (μ): 10^{-6}, nano (n): 10^{-9}, pico (p): 10^{-12}, femto (f): 10^{-15}. atto (a): 10^{-18}, zepto (z): 10^{-21} and yocto (y): 10^{-24}. The tokenisation schemes on the abscissae are sorted according to the average length of tokens in ascending order.

Let us now turn to analysing the effect of masking the corpus. Figure 5 shows similar information to Figure 4 but is specific to the Tatoeba corpus in English. These figures show the comparison between masking and not masking the corpus based on their tokens' frequencies. We found that the number of analogies are significantly higher when we perform masking, both the least and most frequent. This is also true for the analogical density and proportion of analogical sentences. The striking result is how the analogical density improves significantly when we mask the least frequent tokens. In this case, we found that masking the least frequent tokens on sentences tokenised with the sub-word tokenisation scheme increases the analogical density by up to 10^4 times. Analogical density also increases when we masked with the most frequent method even though it is not as much as the least frequent one. Thus, we can observe that the proportion of analogical sentences increases up to 6 times when we masked the least frequent tokens. This shows that masking help increases the analogical density.

Although we do not show the plots for the other languages, previously mentioned phenomena are also observed in all of the other languages.



(**a**) Number of extracted analogical clusters

(**b**) Number of extracted analogies

(**c**) Analogical densities

(**d**) Proportion of sentences appear in analogy

Figure 5. Same as Figure 4 but for the Tatoeba corpora in English with and without masking. There are two masking methods: least frequent (*Tatoeba-least*) and most frequent (*Tatoeba-most*). The tokenisation schemes on the abscissae are sorted according to the length of average tokens in ascending order. Caution: The maximum value of the ordinates are different with Figure 4.

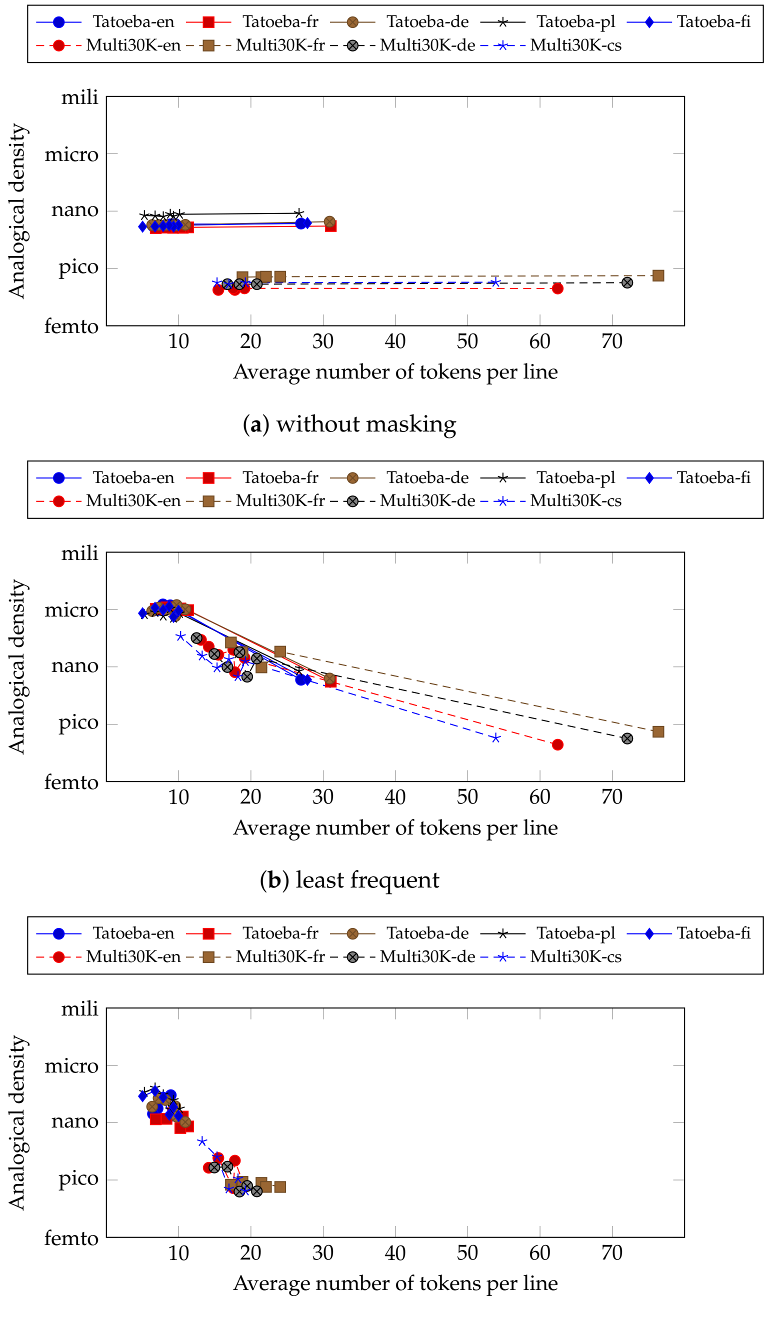
7.2. Impact of Average Length of Sentences on Analogical Density

Figure 6 shows the plot between analogical density against the average number of tokens per line. The number of tokens is calculated based on their respective tokenisation schemes. If the character tokenisation scheme is used, then the number of tokens per line is just the number of characters that appear per line. For sub-word tokenisation, it is the number of sub-word tokens that appear per line. Lastly, it is the number of words for the word tokenisation scheme.

We can immediately observe that analogical densities are different between corpora even when there is overlap on the average number of tokens per line. This also holds when having different languages in the same corpus. This shows that analogical density is particular to the sentences contained in a corpus rather than the average number of tokens per line. Thus, we may confidently conclude that analogical density is influenced more by the type of sentences contained in the corpus rather than the number of tokens inside the sentence.

However, we can observe a stable increase in analogical density for each of the corpora on the situation without masking (Figure 6a). The higher average number of tokens leads to higher analogical densities. This means that we can increase the analogical density of a corpus by increasing the number of tokens. This can be achieved by tokenising the corpus into a more granular one. Particularly, in this paper, we achieved that by decreasing the vocabulary size of the sub-word tokenisation scheme.

When we masked the tokens, both least frequently (Figure 6b) and most frequently (Figure 6c), the analogical density increases as the number of tokens per line rises and then decreases after some time. This is different from the non-masking situation where we observe a stable increase along with the number of tokens per line. Similar to the previous observation, masking the least frequent tokens gives more improvements than masking the most frequent tokens. From this, we conclude that both masking and sub-word tokenisation is an important factor of increasing the analogical density of a given corpus.



(**a**) without masking

(**b**) least frequent

(**c**) most frequent

Figure 6. Analogical density against the average number of tokens per line (respective to their tokenisation schemes) for the Tatoeba and Multi30K corpora in several languages. The three plots are without masking (**a**), with masking for the least frequent (**b**) and the most frequent (**c**).

8. Further Discussion

8.1. Vocabulary Size of Sub-Word Tokenisation

We performed experiments with varying sizes of vocabulary as the parameter for sub-word tokenisation. From the results, we can see higher vocabulary sizes. However, there seems to be a peak before the analogical density drops. We also found that the vocabulary sizes are different for each corpus. As our goal is to maximise the analogical density we determine this parameter automatically.

8.2. Masking Ratio

From the experimental results, the use of masking proved to be an effective way to increase analogical density.

Currently, we mask the tokens in sentences with a ratio of 50% based on their frequency. Same as the previous discussion on vocabulary size, it would be convenient if we can immediately determine the masking ratio or even a different way to mask the sentences in the corpus. Table 10 illustrates the masked sentence under different masking ratio situations.

Table 10. Illustration on the how masking ratio may influence the appearance of a sentence. In this example, we mask the most frequent tokens.

Ratio (%)	Masked Sentence						
0	*the*	*store*	*is*	*closing*	*at*	*7*	*.*
25	...	*store*	*is*	*closing*	*at*	*7*	...
50	...	*store*	...	*closing*	*at*	*7*	...
75	...	*store*	...	*closing*	...	...	...
100	...	...	...	...	...	...	...

8.3. The Level of Analogy: Surface Form and Distributional Semantics

In this work, we only consider analogies on the level of form. In future work, it will be interesting to also confirm the analogy on the level of semantics. In this case, we may use word embeddings, which is a popular approach in distributional semantics.

9. Conclusions

We performed experiments in measuring the analogical density of various corpora in various languages using different tokenisations and masking tokens by their frequencies. To compute these analogical densities we extracted analogical clusters and counted the number of analogies. We also measured the proportion of sentences that appear in at least one analogy. From all our experimental results, we state the three following main findings.

- Corpora with a higher Type–Token Ratio tend to have higher analogical densities.
- We naturally found that the analogical density goes down from the character to word. However, this is not true when tokens are masked based on their frequencies.
- Masking tokens with lower frequencies leads to higher analogical densities.

As a conclusion, in order to increase the analogical density of a corpus, we recommend using the following techniques:

- Use sub-word tokenisation, and vary the size of the vocabulary to maximise the Type–Token Ratio.
- If the task allows for it, mask tokens with lower frequencies and vary the threshold to maximise the Type–Token Ratio again.

For future work, we are also interested in knowing the connection between our proposed measurement and the NLP task. We expect that machine translation tasks performed on a corpus with higher analogical density result in better translation quality. This is very much anticipated for machine translation systems that rely on case-based reasoning, such as EBMT system. Higher analogical density means more chances to *reuse* sentences as facts to perform translation, for example, through analogy. In summary, increasing the analogical density by sub-word tokenisation and masking the tokens with lower frequencies could be performed as a data preprocessing task to assist machine translation systems. Further experiments to measure the correlation between the metrics are needed to investigate this hypothesis.

Author Contributions: Conceptualization, R.F. and Y.L.; methodology, R.F. and Y.L.; software, R.F.; validation, R.F. and Y.L.; formal analysis, R.F. and Y.L.; investigation, R.F. and Y.L.; resources, R.F.; data curation, R.F.; writing—original draft preparation, R.F.; writing—review and editing, R.F. and Y.L.; visualization, R.F.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a JSPS grant, number 18K11447 (Kakenhi Kiban C), entitled "Self-explainable and fast-to-train example-based machine translation".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data, tools and/or program used in this research are all publicly available for download through lepage-lab.ips.waseda.ac.jp/en/projects/kakenhi-15k00317/ (accessed on 20 September 2020) or the links provided in the body of the paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Hathout, N. Acquisition of morphological families and derivational series from a machine readable dictionary. *arXiv* **2009**, arXiv:0905.1609.
2. Lavallée, J.F.; Langlais, P. Morphological acquisition by formal analogy. In *Morpho Challenge 2009*; Knowledge 4 All Foundation Ltd.: Surrey, UK, 2009.
3. Blevins, J.P.; Blevins, J. (Eds.) Analogy in Grammar: Form and Acquisition. Oxford Scholarship Online. 2009. Available online: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199547548.001.0001/acprof-9780199547548 (accessed on 25 July 2021).
4. Fam, R.; Lepage, Y. A study of the saturation of analogical grids agnostically extracted from texts. In Proceedings of the Computational Analogy Workshop at the 25th International Conference on Case-Based Reasoning (ICCBR-CA-17), Trondheim, Norway, 26–28 June 2017; pp. 11–20. Available online: http://ceur-ws.org/Vol-2028/paper1.pdf (accessed on 25 July 2021).
5. Wang, W.; Fam, R.; Bao, F.; Lepage, Y.; Gao, G. Neural Morphological Segmentation Model for Mongolian. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN-2019), Budapest, Hungary, 14–19 July 2019; pp. 1–7.
6. Langlais, P.; Patry, A. 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-07), Prague, Czech Republic, 28–30 June 2007; pp. 877–886. Available online: https://aclanthology.org/D07-1092 (accessed on 25 July 2021).
7. Lindén, K. Entry Generation by Analogy—Encoding New Words for Morphological Lexicons. *North. Eur. J. Lang. Technol.* **2009**, *1*, 1–25. [CrossRef]
8. Fam, R.; Purwarianti, A.; Lepage, Y. Plausibility of word forms generated from analogical grids in Indonesian. In Proceedings of the 16th International Conference on Computer Applications (ICCA-18), Beirut, Lebanon, 25–26 July 2018; pp. 179–184.
9. Hathout, N.; Namer, F. Automatic Construction and Validation of French Large Lexical Resources: Reuse of Verb Theoretical Linguistic Descriptions. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.549.5396&rep=rep1&type=pdf (accessed on 25 July 2021).
10. Hathout, N. Acquistion 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, Manchester, UK, 24 August 2008; pp. 1–8. Available online: https://aclanthology.org/W08-2001 (accessed on 25 July 2021).
11. Lepage, Y.; Denoual, E. Purest ever example-based machine translation: Detailed presentation and assessment. *Mach. Transl.* **2005**, *19*, 251–282. [CrossRef]
12. Takezawa, T.; Sumita, E.; Sugaya, F.; Yamamoto, H.; Yamamoto, S. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Spain, 29–31 May 2002. Available online: http://www.lrec-conf.org/proceedings/lrec2002/pdf/305.pdf (accessed on 25 July 2021).

13. Lepage, Y. Lower and Higher Estimates of the Number of "True Analogies" between Sentences Contained in a Large Multilingual Corpus. Available online: https://aclanthology.org/C04-1106.pdf (accessed on 25 July 2021).
14. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: The Tenth Machine Translation Summit*; AAMT: Phuket, Thailand, 2005; pp. 79–86.
15. Lepage, Y. Solving Analogies on Words: An Algorithm. In Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998), Montreal, QC, Canada, 10–14 August 1998; Volume 1, pp. 728–734.
16. Stroppa, N.; Yvon, F. An Analogical Learner for Morphological Analysis. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, MI, USA, 29–30 June 2005; pp. 120–127.
17. Langlais, P.; Yvon, F. Scaling up Analogical Learning. In Proceedings of the Coling 2008: Companion Volume: Posters, Manchester, UK, 18–22 August 2008; pp. 51–54.
18. Beesley, K.R. Consonant Spreading in Arabic Stems. COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics. Available online: https://aclanthology.org/C98-1018.pdf (accessed on 25 July 2021).
19. Wintner, S. Chapter Morphological Processing of Semitic Languages. In *Natural Language Processing of Semitic Languages*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 43–66.
20. Gil, D. From Repetition to Reduplication in Riau Indonesian. In *Studies on Reduplication*; De Gruyter: Berlin, Germany; pp. 31–64.
21. Lepage, Y. Analogies Between Binary Images: Application to Chinese Characters. In *Computational Approaches to Analogical Reasoning: Current Trends*; Prade, H., Richard, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 25–57.
22. Elliott, D.; Frank, S.; Sima'an, K.; Specia, L. Multi30K: Multilingual English-German Image Descriptions. In Proceedings of the 5th Workshop on Vision and Language, Berlin, Germany, 7–12 August 2016; pp. 70–74.
23. Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; Specia, L. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, Copenhagen, Denmark, 7–8 September 2017; pp. 215–233.
24. Barrault, L.; Bougares, F.; Specia, L.; Lala, C.; Elliott, D.; Frank, S. Findings of the Third Shared Task on Multimodal Machine Translation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 304–323.
25. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
26. Kudo, T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 66–75. Available online: https://aclanthology.org/P18-1007 (accessed on 25 July 2021).
27. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725. Available online: https://aclanthology.org/P16-1162 (accessed on 25 July 2021).
28. Provilkov, I.; Emelianenko, D.; Voita, E. BPE-Dropout: Simple and Effective Subword Regularization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 5–10 July 2020; pp. 1882–1892. Available online: https://aclanthology.org/2020.acl-main.170 (accessed on 25 July 2021).
29. Koehn, P.; Knight, K. Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, Austin, TX, USA, 30 July–3 August 2000; pp. 711–715.