Original Article

How Skewed is the World for a Language?: A Computational Approach to the Semantic Distribution of Languages

Takuto Nakayama

Faculty of Health Sciences, Kyorin University (part-time).

5-4-1 Shimorenjaku, Mitaka, Tokyo. 181-0013.

Email: tnakayama.a5ling@gmail.com

Abstract

This paper aims to illustrate the skewness among distributions of semantic language categories and whether such distributions are common. As the methodology, this research used an automatic semantic tagging tool, the USAS semantic tagger, for eight languages: Chinese, Dutch, English, Finnish, French, Italian, Portuguese, and Spanish. Shannon entropies and three-dimensional reduction methods, MDS with Jensen-Shannon divergences and tSNE, were also conducted to measure similarities of distributions of semantic tags in each language. As a result, it was suggested that the patterns of semantic tags fall into a tiny space, while they are unique to each language within the space. Moreover, it was suggested that the rank frequency distributions of the semantic categories of languages that are typologically similar or closely related to the same language family could be close.

Keywords: Semantic distribution, UCREL Semantic Analysis System, MDS, tSNE

1 Introduction

Languages are used to represent various meanings worldwide. Technically, the connection between form and meaning—constituting the form-meaning pairing—enables the categorization of languages at the semantic poles into numerous groups, including abstract concepts, physical features, and art- and emotion-related constructs. In such groups, it can be observed that the number of expressions categorized into one group is much higher than that categorized into another, in which case the language presenting a similar distribution of meanings would perceive the world as "skewed." As such, this paper aims to illustrate the skewness among distributions of semantic language categories and whether such distributions are common.

Section 2 will introduce previous studies on equi-complexity to explain the motivation behind this research, including reasons for studying semantic distributions. Section 3 will explain the methodology, which consists of three measurements : Shannon entropy and Multi-dimensional Scaling (MDS) by Jensen-Shannon divergence, and t-Distributed Stochastic Neighbor (tSNE). Finally, Section 4 and 5 will describe the results and discussion, respectively.

2 Previous Studies

2.1 Equi-complexity

The notion that all languages are equally complex is generally called equi-complexity. When comparing two languages, a relationship between certain language facets is identified. For example, language A might have a more complex inflectional system than language B, whereas language B might have a larger vocabulary than language A. It is suggested that even if one language facet is more complex than all others, there will remain at least one language facet that is simpler than the others, suggesting that the overall complexity of languages is equal.

The concept of equi-complexity was first introduced nearly 70 years ago, when Hockett (1958) asserted, "[I]mpressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do" (Hockett, 1958, p. 180). As a background to equi-complexity, it can be argued that linguists, especially those with an anthropological background, identified the concept as the antithesis to the idea based on the imperialist policy that languages used by uncivilized groups are simpler. Everett (2005) claimed, "No one should draw the conclusion from the paper that the Pirahã language is in any way 'primitives.' It has the most complex verbal morphology I am aware of. And a strikingly complex prosodic system" (Everett, 2005, p. 62).

Although equi-complexity has been well-understood among linguists for a long time, whether the concept is true has not yet been corroborated. Since the end of the 20th century, the development of computer technology has made it possible to assess a larger amount of data than what human beings could previously manage. Now, equi-complexity is one of the most intriguing topics in linguistic research, especially in computational linguistics.

2.2 Formalism in Computational Linguistics

Computational linguistics is a subfield of linguistics in which equi-complexity is one of the major topics calculated in many ways. For example, Ehret and Szmrecsanyí (2016) proposed calculating Kolmogorov complexity by compressing a file into a zipped file to measure overall linguistic complexity, where Kolmogorov complexity is defined as "a string/text as the length of the shortest possible description of that string/text." (Ehret & Szmrecsanyí, 2016, p. 72). Thus, the difference between the original file size and its compressed file size represents the Kolmogorov complexity of the file. This proposed method is quite useful, as it can be applied to a vast number of files automatically, enabling the approximation of a language's overall complexity. However, as this method, it considers only the formal information of a file, not the semantic facets.

This attitude toward semantics is due to the formalism in computational linguistics, which focuses on criteria that can identify whether a stream is meaningful. Zipf (1935, p. 187) asserted:

> If a Martian scientist sitting before his radio in Mars accidentally received from Earth the broadcast of an extensive speech which he recorded perfectly through the perfection of Martian apparatus and studied at his leisure, what criteria would he have to determine whether the reception represented the effect of animate process on Earth, or merely the latest thunderstorm on Earth? It seems that the only criteria would be the arrangement of occurrences of the elements, and the only clue to the animate origin would be this: the arrangement of the occurrences would be neither of rigidly fixed regularity such as frequently found in wave emissions of purely physical origin nor yet a completely random scattering of the same.

As Zipf argued, computational linguists are interested in distinguishing the natural language of human beings from other codes, such as animal sounds and random streams of letters, while observing the features of the formal complexity of a code. In addition, information theory, which provides methods and measurements, such as Shannon entropy and Kolmogorov complexity, to determine linguistic complexity, does not consider semantics as closely. Shannon (1948, p. 379) claimed:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. (Shannon, 1948, p. 379)

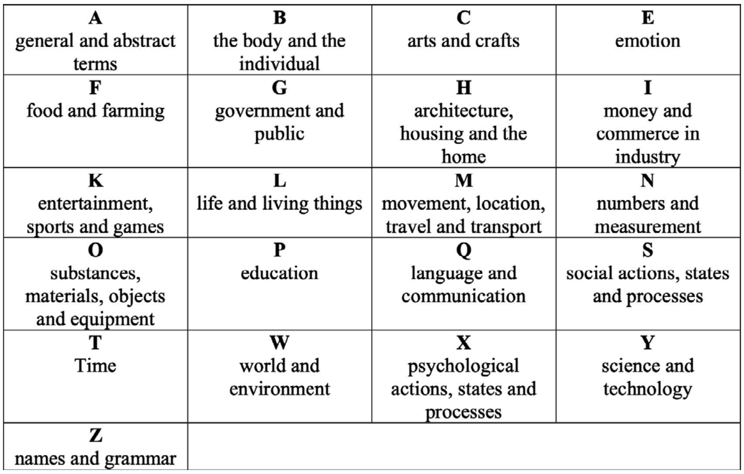
Thus, little research mentions the semantic facets of a language, including Bentz (2023), for instance.

However, this paper argues semantics is as important as the formal facets of a language, as a language consists of a set of form–meaning pairings, as mentioned in Section 1. Therefore, this research examines to approach the complexity of language semantics.

3 Methodology

3. 1 The UCREL Semantic Analysis System

This study used the University Centre for Computer Corpus Research on Language (UCREL) semantic analysis system (USAS)[1] (Rayson et al., 2004) to annotate each word in the data. According to the website, USAS "has been designed and used across several research projects and this page collects together various pointers to those projects and publications produced since 1990" (Lancaster University, 2008). The tagset consists of 21 major categories and several minor categories, some of which have subcategories, where the former is represented by capital letters and the latter by numbers.



A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Figure 1 Tag categories in the USAS semantic tagger (Archer et al., 2002, p. 2)

The following examples are an annotated text from the Gospel of Matthew (Table 1):

Table 1 Example of annotation

The	book	of	the	generation	of	Jesus	Christ	,
Z5	Q4	Z5	Z5	T1	Z5	Z4	Z4	PUNCT
the	son	of	Abraham	.				
Z5	Z2	Z2	Z2	PUNCT				

The second word of the sentence, "book," is given the tag, Q4, in which Q refers to a major category, "Language and communication," and 4 to a minor one, "The media." On the other hand,

[1] https://ucrel.lancs.ac.uk/usas/

the first word, “the,” has Z5, which means “Grammatical bin” in the major category “Names and grammar.”

The reasons for using this semantic tagging system are that it allows one to annotate each word automatically, and it can be applied to multiple languages. USAS has been developed in several programming languages, including C and JavaScript, but in this study, a module developed in Python by Scott Piao and Andrew Moore, called the Python Multilingual UCREL semantic analysis system (PyMusas), was used, as it can apply an automatic annotation to a vast amount of data in a short time, enabling one to compare multiple languages. USAS has also expanded its areas of application; in the latest version, 10 languages are available: Chinese, Dutch, English, Finnish, French, Indonesian, Italian, Portuguese, Spanish and Welsh. However, this research used eight of them, except Indonesian and Welsh because it was no longer available to access a tagging system for preprocessing raw texts, called CyTag toolkit, wrapped in a docker container.

3.2 Shannon Entropy

To compare the distributions of semantic tags in the texts, Shannon entropy (Shannon, 1948) is introduced, the value of which refers to the average of uncertainty concerning which “alphabet” will appear next, where “alphabet” refers to an entity building an informational stream.

In other words, it is the degree to which the distribution is skewed into one tag, or every probability is near equal. Shannon entropy H is defined as follows:

$$H = -\sum_{i} p(i)\log_2 p(i),$$

where $p(i)$ refers to the probability of i occurring, and H is always positive. In addition, the lower the value of H is, the more skewed the distribution is; the higher H is, the less skewed or more similar each value in a distribution is, and vice versa.

3.3 Jensen-Shannon Divergence

As another measure of the variance in probability distributions, JS divergence measures the distance between the divergence in one text and that in another, called Jensen-Shannon (JS) divergence (Li, 1991). However, before introducing JS divergence, we must define another measurement must be defined, Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), which illustrates the difference between the probability distributions P and Q, defined as follows:

$$D_{KL}(P \parallel Q) = p(i)\log_2 \frac{p(i)}{q(i)},$$

where $p(i)$ and $q(i)$ refer to the probabilities of each phenomenon.

Furthermore, JS divergence is defined along with KL divergence as follows:

$$D_{JS}(P \parallel Q) = \frac{1}{2}(D_{KL}(P \parallel R) + D_{KL}(Q \parallel R)),$$

where $R(i) = \frac{1}{2}(p(i) + q(i))$.

Both D_{KL} and D_{JS} are always positive, and 0 means that the probability distributions are the same. Furthermore, the difference between KL divergence and JS divergence is that $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$, whereas $D_{JS}(P \parallel Q) = D_{JS}(Q \parallel P)$. Thus, KL divergence is not exactly a "distance" between probability distributions in that the measurement is asymmetrical. In other words, JS divergence arranges KL divergence arranged to be a distance with a symmetrical feature. Thus, this study uses JS divergence as a measure of differences among distributions of semantic tags in texts.

3.4 Dataset and Procedure

Table 2 Overview of corpora

Language	Corpus	Size (words)
Chinese	zlTenTen17	13,531,331,169
Dutch	nlTenTen20	5,890,009,964
English	enTenTen20	52,268,286,493
Finnish	fiTenTen	1,404,100,049
French	frTenTen	15,115,914,647
Italian	itTenTen20	12,451,734,885
Portuguese	prTenTen20	12,578,775,252
Spanish	esTenTen18	16,951,839,897

The dataset comprises eight languages: Chinese, Dutch, English, Finish, French, Italian, Portuguese and Spanish. Ten sample files were created for each language, each almost 1MB (or 1,000,000 characters, as 1MB is the upper limit of PyMusas' capability.) The sentences were randomly obtained from Sketch Engine, a hub interface that provides many kinds of corpora of multiple languages. The reason these languages were chosen is that they can be derived from Sketch Engine and can be tagged by the USAS Semantic tagger.

Next, the data given by the procedure above were tagged by the USAS semantic tagger. During this phase, PyMusas was processed, after which the tag frequencies of the tags were arranged into the rankings, and the relative frequencies and probabilities of appearance were calculated. Based on these results of the data arrangement, the similarities among the semantic tag distributions of the texts were calculated with two methods, JS divergence plotted by multi-dimensional scaling and vectors by t-Distributed Stochastic Neighbor (tSNE), where all calculations are automatically derived using Python[2]. These methods enable us to plot the results of the analysis in a lower dimensional space. In the next section, the analysis results will be embedded in a three-dimensional space to visualize the similarities among languages.

4 Results

4.1 Annotated Data

To provide an overview of the annotating process, Table 3 illustrates the ratios of major semantic categories of USAS for the eight languages. The values represent the summed values of the 10 files of each language. Scratching the surface, the distribution of the ratios seems to be quite similar among all the languages. The following subsections will scrutinize the annotated data in detail, focusing on the minor semantic categories.

In addition, to check how appropriately the taggers for each language in USAS annotate the texts, an evaluation test was conducted. In the test, parallel texts translated into each language, the Gospel of Matthew from Bible Parallel Corpus[3], were annotated. As the benchmark, on the one hand, the tag for the English text was supposed to be the most appropriate ; the other languages were calculated in terms of their distance from the English text based on the KL divergence of their probability distributions of the tags between English and the other languages texts. As Table 4 shows, the distributions of semantic tags in Dutch, Finnish, Italian, and Portuguese texts, which have lower KL divergence, are quite similar to the ones in English, indicating that these four languages were tagged by USAS adequately. In the case of Chinese, French and Spanish, the KL divergences are relatively high, meaning that they are not similar to the distribution of semantic tags in the English texts. This seems to be because many words were categorized into unknowns, represented by the tag Z99 in the USAS tag system. Thus, in the following analysis, all words tagged as “Z” will be ignored to make the result more adequate.

[2] All Python codes this research used are uploaded in GitHub (https://github.com/takuto-nakayama/semantics_distribution)

[3] https://christos-c.com/bible/

Table 3 Ratios of tags for each language

Tag	Chinese	Dutch	English	Finnish	French	Italian	Portuguese	Spanish
A	0.232	0.231	0.230	0.227	0.224	0.220	0.214	0.206
B	0.019	0.019	0.019	0.019	0.018	0.018	0.017	0.016
C	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
E	0.011	0.011	0.010	0.010	0.010	0.010	0.009	0.009
F	0.010	0.010	0.010	0.010	0.009	0.009	0.008	0.008
G	0.011	0.011	0.011	0.011	0.010	0.010	0.010	0.009
H	0.010	0.010	0.010	0.010	0.010	0.009	0.009	0.008
I	0.022	0.022	0.022	0.021	0.021	0.020	0.019	0.018
K	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005
L	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003
M	0.032	0.032	0.033	0.034	0.035	0.037	0.039	0.042
N	0.104	0.103	0.102	0.101	0.099	0.096	0.092	0.087
O	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012
P	0.008	0.008	0.008	0.007	0.007	0.007	0.006	0.006
Q	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015
S	0.024	0.024	0.024	0.024	0.024	0.024	0.024	0.023
T	0.019	0.019	0.019	0.019	0.020	0.020	0.021	0.023
W	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
X	0.013	0.013	0.013	0.013	0.014	0.014	0.015	0.016
Y	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Z	0.441	0.442	0.445	0.449	0.455	0.463	0.473	0.487

Table 4 KL divergence for the English texts

Language	Chinese	Dutch	Finnish	French	Italian	Portuguese	Spanish
KL divergence	0.553	0.271	0.345	0.812	0.292	0.316	0.489

4.2 Entropy

Table 5 Entropy of each file

File ID	Chinese	Dutch	English	Finnish	French	Italian	Portuguese	Spanish
1	5.315	5.425	5.647	5.715	5.649	5.879	5.497	5.822
2	5.323	5.423	5.652	5.706	5.644	5.880	5.502	5.821
3	5.332	5.417	5.655	5.702	5.650	5.876	5.508	5.825
4	5.336	5.416	5.656	5.706	5.647	5.876	5.507	5.824
5	5.336	5.414	5.660	5.706	5.648	5.877	5.509	5.825
6	5.338	5.413	5.663	5.706	5.650	5.876	5.510	5.823
7	5.337	5.411	5.664	5.707	5.652	5.878	5.512	5.824
8	5.338	5.411	5.665	5.708	5.651	5.878	5.512	5.823
9	5.338	5.413	5.665	5.707	5.652	5.878	5.512	5.823
10	5.338	5.412	5.664	5.707	5.653	5.878	5.511	5.823
average	5.333	5.415	5.659	5.707	5.650	5.877	5.508	5.824
standard deviation	0.008	0.005	0.006	0.003	0.003	0.001	0.005	0.003

Entropies of each language are shown in Table 5, though it is suggested that Shannon entropy cannot distinguish languages because the differences between entropies are too small. If an entropy is high, the probability distribution is less skewed, while the opposite is also true. Looking at the table, Chinese has the lowest entropy and Spanish the highest, between which the others are found.

4. 3 JS Divergence

Figure 2 visualizes the distance among each text file based on a set of the JS divergences. Each dot corresponds to each file, and the distances between the dots represent JS divergences. Thus, x–axis, y–axis and z–axis values refer simply to the distances between dots.

As shown in Figure 2, all the languages are grouped into clusters in which the dots are plotted almost at the same point. In addition, Chinese texts are plotted in one corner of the 3D space, while the other languages are placed relatively close to each other. First, the figure suggests that all languages have their unique semantic distributions. Second, under such a condition, Chinese differs from the other languages that are relatively close to each other.

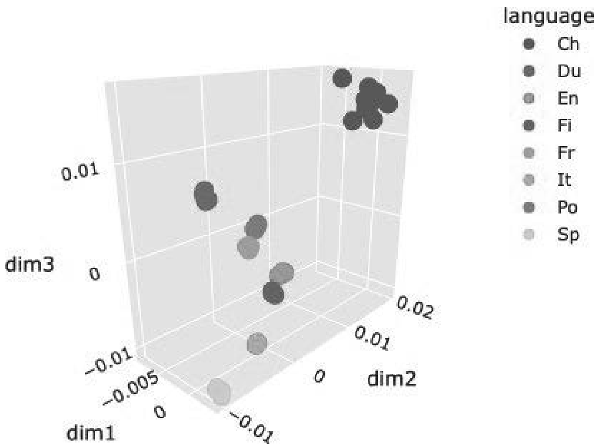


Figure 2 3D–Scatter Plot of JS divergence

On the other hand, Figure 3 visualizes the result of tSNE for the probability distributions of the semantic tags in each language. This tSNE process is under that the perplexity is 5. As shown in Figure 3, each language creates its clusters, meaning that, as Figure 2 shows, each language has a unique distribution pattern of the semantics.

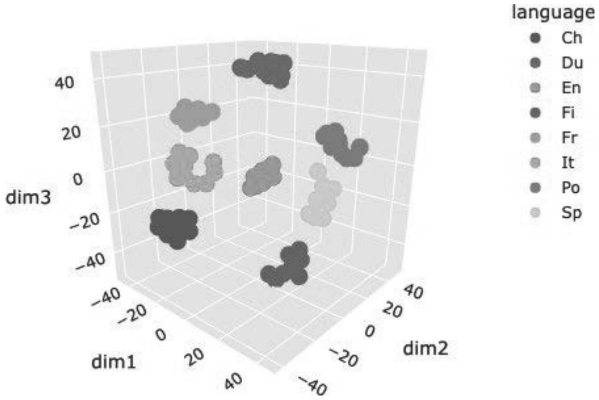


Figure 3 3D-Scatter Plot in tSNE

5 Discussion

According to the result from calculating the Shannon entropies, there is no large gap among each file of each language, implying that each language has a similar distribution of semantic tags. The JS divergence in MDS (Figure 2) shows that, while the Chinese texts are different from the other languages, they are relatively close to each other. However, all dots are placed in a small three-dimensional space, in which the ranges of all axes are all less than 0.03. Thus, distributions of semantics in a language seem not to vary but fall into a tiny space, while they are grouped into a cluster within a space, illustrating the uniqueness of the distributions. What must be kept in mind is that the distributions mentioned above do not focus on the most frequent semantic tags but on the rank frequency distributions of the semantic tags. In other words, there are no similarities in the kinds of meanings likely to appear most often, but there are similarities in how vocabulary is categorized into different semantic tags based on the proportion that falls into each category. It is natural for a language to have a unique semantic category that is the most common; for example, it is easily expected that languages in cold environments (e.g., Finland) will have a more diverse vocabulary to convey something unique that is not used as often as somewhere warm or hot. Even in such a case where the number of words belonging to one semantic category is significantly high, suggesting that the number of words divided into another category is low, one value balances out the other.

On the other hand, the result of tSNE seems to indicate another aspect of the distribution of semantics.

In Figure 3, the vectors of probability distributions of semantic tags are plotted. That the vectors of each language create their own clusters demonstrates the uniqueness of semantic distribution patterns in a language. Furthermore, the clusters appear to be arranged with English at the center, and other languages are positioned around it in all directions, both horizontally and vertically, except Chinese. This implies that English and other languages roughly share values in two dimensions. In particular, concerning the third dimension represented by "dim3" in Figure 3, Romance languages share their values, while, on one end, Dutch, a Germanic language, is placed, on the one end of this dimension and, on the other end, Finnish, a Finno-Ugric language, is on the other end. This partially suggests typological similarity among languages, with Romance languages showing shared values and Dutch and Finnish, representing Germanic and Finno-Ugric languages, respectively, positioned at opposite ends. However, the reason for "partially" is because not all the four Romance languages share all values on the dimensions. Italian and Portuguese share the values on the first and third dimensions but not the second ; French and Spanish share the second and third but not the first. What is interesting might be that the Chinese cluster does not share their values in all dimensions. This also suggests that language family might be one of the major factors in determining its semantic distribution pattern. On top of that, there is no diagonal relation between each language in Figure 3. This suggests that, at least, these eight languages do not have any correlation among each dimension.

Because Chinese is distinct from the other languages, it might be that the observed patterns of semantic distributions can be shared in languages that are typologically similar or closely related to the same language family. However, this suggestion would be limited unless a language in another language family, say Asian languages, is also scrutinized.

To sum up, the Shannon entropies and JS divergences, which were quite similar among each language, suggest that the equi-complexity of language might be true, while each language has its own probability distributions of semantic tags, from the result of tSNE, where each language creates their unique clusters.

6 Conclusion

This research used an automatic semantic tagging tool, the USAS semantic tagger, for eight languages : Chinese, Dutch, English, Finnish, French, Italian, Portuguese and Spanish. Shannon entropies and three-dimensional reduction methods, MDS with JS divergences and tSNE, were also conducted to measure the similarities of the distributions of semantic tags in each language. As a result, there were no large gaps among the Shannon entropies of each file, whereas the distances among languages showed some differences, especially between Chinese and the other seven languages. From the result, the patterns of semantic tags fall into a tiny space, while they are unique to each language within the space. Moreover, it was suggested that the rank frequency distributions of the semantic categories of languages that are typologically similar or closely related to the same language family could be close. Conversely, because Chinese is the only non-Indo-European putout language dealt with, whether the behavior of the distribution described above is always true has not been verified, which is a major limitation of this research.

References

Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system.*

Bentz, C. (2023). Beyond words: Lower and upper bounds on the entropy of subword units in diverse languages. *16th International Cognitive Linguistics Conference*, Düsseldorf, Germany.

Ehret, K., & Szmrecsanyi, B. (2016). An information theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity, isolation, and variation* (vol. 57), pp. 71–94.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology, 46*(4), 621–646. https://doi.org/10.1086/431525

Hockett, C. F. (1958). *A Course in modern linguistics.* Macmillan.

Kortmann, B., & Schröter, V. (2020). *Linguistic complexity.* https://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0254.xml.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Lancaster University (2008) *UCREL semantic analysis system* https://ucrel.lancs.ac.uk/usas/

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151.

Rayson, P., Archer, D., Piao, S. L., & McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic labeling for NLP Tasks in association with 4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, 7–12.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology.* Houghton Mifflin.