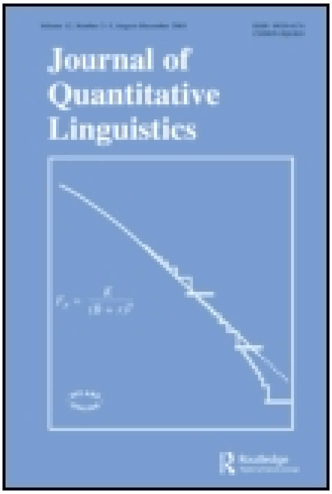

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Identifying universals of text translation*

Ido Kanter [a] , Haggai Kfir [a] , Brenda Malkiel [b] & Miriam Shlesinger [b]

[a] Department of Physics , Bar-Ilan University , Ramat-Gan, Israel

[b] Department of Translation and Interpreting Studies , Bar-Ilan University , Ramat-Gan, Israel

Published online: 16 Feb 2007.

Journal of Quantitative Linguistics
2006, Volume 13, Number 1, pp. 35–43
DOI: 10.1080/09296170500500983

Identifying Universals of Text Translation*

Ido Kanter[1], Haggai Kfir[1], Brenda Malkiel[2] and Miriam Shlesinger[2]
[1]Department of Physics, Bar-Ilan University, Ramat-Gan, Israel; [2]Department of Translation and Interpreting Studies, Bar-Ilan University, Ramat-Gan, Israel

ABSTRACT

Straightforward quantitative analyses of authentic texts have allowed linguists and translation scholars to discern patterns in individual languages as well as features which set translations apart from originals (Baker, 1993; Chesterman, 2004). A language can also be studied statistically, an approach epitomized by the application of Zipf's Law (Zipf, 1949), which states that word-frequency distributions follow an almost identical curve regardless of language. To date, no universal behaviour governing the joint probability distribution of words in two or more languages has been either proposed or observed. This study identifies new universals which characterize the mutual overlaps between a corpus of original English and three corpora of translated English. Specifically, it suggests a remarkable similarity in (a) the number of types unique to each translated corpus, and (b) the number of types common to the original-English corpus and each of the translated corpora. We argue that these universal behaviours can be used both to determine the ontological status of an unidentified language (whether it is an original or a translation) and to identify the source language of a translation.

Corpus-based linguistics and corpus-based translation studies have used corpora – large bodies of authentic texts in machine-readable format – to arrive at generalizations about language in use rather than language systems in the abstract, and to discern universals of translation – features that are independent of any particular language pair, text type, translator, or historical period (Chesterman, 2004). Proposed universals of translation include lexical, syntactic and stylistic simplification (Blum-Kulka & Levenston, 1983; Laviosa-Braithwaite, 1997); a high degree of

*Address correspondence to: Miriam Shlesinger, Department of Translation and Interpreting Studies, Bar-Ilan University, Ramat-Gan, 52900 Israel.
E-mail: shlesm@mail.biu.ac.il

explicitness (Blum-Kulka, 1986); unusual patterns of co-occurrence (Nilsson, 2002; Jantunen, 2004); and over- or under-representation of typical features and lexical items (Vanderauwera, 1985; Kenny, 1998); Tirkkonen-Condit, 2004).

A language or corpus may also be studied statistically, on the assumption that statistical analysis will reveal certain global patterns, such as word-frequency distributions. This approach is epitomized in the application of Zipf's Law, which states that the frequency of a particular event – e.g. of a specific word – (P) as a function of its rank (m) is a power-law function $P_m \sim 1/m^{\alpha}$, with the exponent α close to unity (Zipf, 1949; Gell-Mann, 1994; Kanter & Kessler, 1995; Mantegna et al., 1995; Halibard & Kanter, 1998; Hatzigeorgiu et al., 2001; Choi, 2000) (Fig. 1a). Zipf's Law was later found to apply to other domains as well, including urban populations (Gell-Mann, 1994) and biological sequences (Mantegna et al., 1995; Halibard & Kanter, 1998).

The remarkable feature of Zipf's Law is its ability to predict the shape of a language, regardless of language family or language size. Every natural language examined has been found to follow Zipf's Law, including the three that figure as source languages for the corpora discussed here: Greek (Hatzigeorgiu et al., 2001), Hebrew (Kanter & Kessler, 1995) and Korean (Choi, 2000).

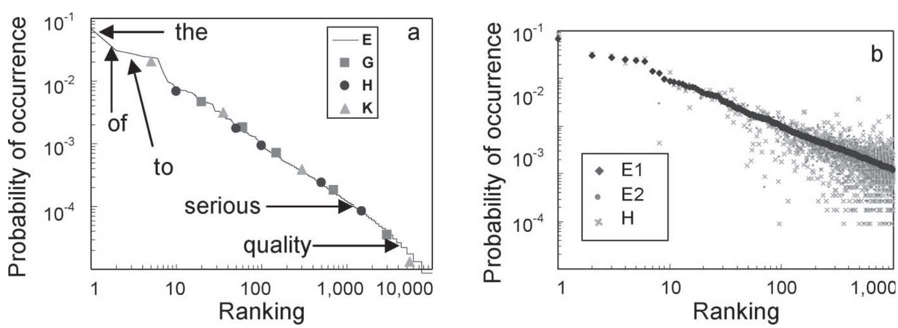

Fig. 1. Zipf's Law for word frequency distributions. (a) The types in a 230,000-word original-English corpus [E] are ordered according to their probability of occurrence. Plotted on a log-log scale, the curve is approximated by a straight line with a slope of -1. The same behaviour is observed in the corpora translated from Greek [G], Hebrew [H] and Korean [K] (since the lines are extremely close, only a few representative words are marked for each corpus). (b) The probability of occurrence of words in a 115,000-word original-English corpus [E1], a 115,000-word corpus of English translated from Hebrew [H] and a second original-English corpus [E2], plotted on a log-log scale, according to the ranking of [E1].

The data for the present study consists of four 230,000-word corpora, one of original English prose and three of English translated from other languages. The data was downloaded from the on-line edition of the *International Herald Tribune* (IHT) and three on-line local supplements to the IHT: the *Kathimerini* English Edition (translated from Greek), *Ha'aretz* (translated from Hebrew), and the *JoongAng Daily* (translated from Korean). Each of the translated corpora represents the collective output of teams of professional translators. The study was limited to English-language editions of the IHT in order to guarantee a high degree of comparability among the corpora.

The composition of all four corpora was the same: news (80,000 words), arts and leisure (50,000), business and finance (50,000), and opinion (50,000). Each of the four newspapers offers different sections; news, arts and leisure, business and finance, and opinion were the only sections common to all four. We refer to the four corpora as the 'English' [E] corpus, the 'from-Greek' [G] corpus, the 'from-Hebrew' [H] corpus, and the 'from-Korean' [K] corpus.

Part of our analysis focuses on data sets containing either one half (115,000 words) or one-quarter of each corpus (57,500 words). These data sets contain the same proportion of news to arts, business, and opinion as the full corpora. We used a lexical analysis software, WordSmith Tools (http://www.lexically.net/wordsmith) to analyse the corpora and determine word frequencies and rankings. Proper nouns were counted as words and the data was not lemmatized.

When viewed individually, each translated corpus appears to follow Zipf's Law, though containing a different set of types (different words) and ranking them differently. The highest-frequency types appear in all three translated corpora examined, with a similar ranking; slightly less frequent words typically appear in all three translated corpora but differ somewhat in ranking; infrequent words do not necessarily appear in all three, and have a markedly different ranking in each.

Notwithstanding its role in describing the universal properties of individual languages, however, Zipf's Law is a rather blunt tool when it comes to comparing languages, since differences are masked by the highest frequency words (often function words) that dominate all natural texts, such as 'the', 'of' and 'to' in the case of English. To date, no universal governing the joint probability distribution of words in two or more languages has been either proposed or observed. A universal would be all the more useful if it were able to point to differences between original texts and translations. The present paper identifies two such universals, and proposes a means of discerning statistical differences between languages, as opposed to those stemming from the subject matter typical of a particular language or other culture-specific factors.

Both original and translated corpora follow Zipf's Law, and their word-frequency distributions follow an almost identical curve (Fig. 1(a)). Typically, we find greater fluctuation between original English and a translated corpus than between two original-English corpora (Fig. 1(b)), but Zipf's Law does not seem capable of pinpointing differences between translations and originals or of determining the source language. To do so, we propose an alternative approach, using Venn diagrams, as shown in Figure 2.

The words in the three comparable corpora were divided into seven groups: the 'core', consisting of words common to all three corpora; three groups comprising words common to two corpora; and three comprising words unique to each specific corpus. The size of each subgroup can be measured in terms of its types, and the analysis is based on comparing the size of the subgroups, thereby bypassing the need to focus on specific lexical items.

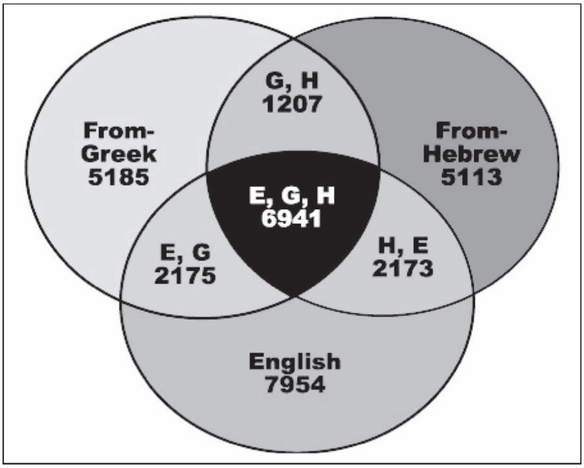

Fig. 2. Venn diagram of the relations between three 230,000-word corpora: original English and translations from Hebrew and Greek. Some words appear in all three languages; others appear in only one or two. The numbers represent the number of types (different words) in each group.

As shown in Figure 3(a), the relationships between the original-English [E] corpus and the ones translated from Greek [G] and from Hebrew [H] are similar: in terms of types, the overlap between [E] and [G] is almost identical to that between [E] and [H] (2175 types and 2173, respectively), and the number of words unique to [G] is very close to the number of words unique to [H] (5185 and 5113, respectively). When we substitute a corpus of translations from Korean [K] for [H], moreover, the pattern is preserved (Fig. 3(b)). Extending this method to a larger number of corpora is a straightforward operation.

This observation points to the following two patterns: (1) the number of words unique to a given translated corpus is similar to that contained by any other matched translated corpus; (2) the number of words unique to an original corpus in a given language and a matched corpus translated into that language is independent of the source language.

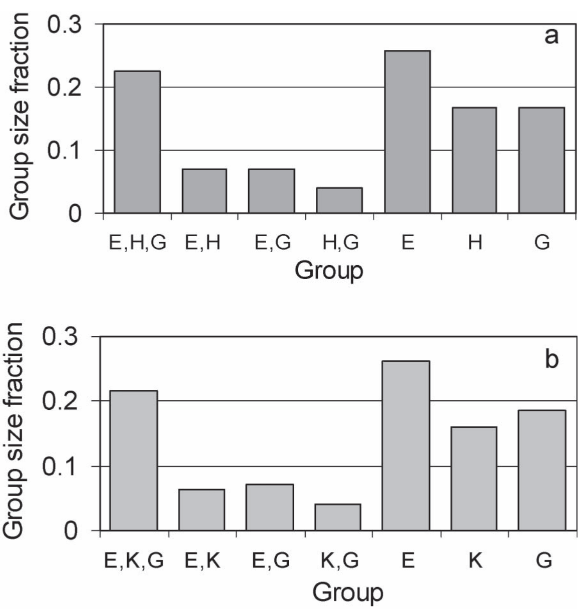

Fig. 3. The number of different words per group, displayed as the fraction of the 230,000-word corpus. Together the four corpora contain 30,754 different types. (a) Original English, English translated from Greek and English translated from Hebrew. (b) Original English, English translated from Greek and English translated from Korean. Note the similarity between 3a and 3b.

A remarkable similarity was also observed with regard to tokens (the total number of appearances of the types) in a subgroup. There are 13,200 tokens unique to [G] and 13,153 unique to [H]. The 2175 types that [E] shares with [G] and the 2173 types that [E] shares with [H] account for 6142 and 6125 tokens, respectively.

One may approach the subject from the reverse direction as well: Can we predict whether an unidentified corpus [U] is original or translated and can we determine its source language? Assuming we know that a given database is either original English or [G] or [H], we could adopt the following naïve procedure:

- Calculate the Venn diagrams for [U]-[G]-[H], [E]-[U]-[H] and [E]-[G]-[U], using the known-source training databases for [E], [H] and [G].
- Identify the language of [U] based on its similarity to the Venn diagrams, by comparing the fraction and/or the content of the seven different sections or the overlaps of the sections.

Figure 4(a) presents the dependence of the seven sections as a function of the size of the databases – 57,500, 115,000 and 230,000 words – and points to the similarity between any two translated corpora and the relationship between these corpora and [E]. In contrast to Zipf's Law, which is predicated on the frequency of a very large number of types, the method proposed here is based on the analysis of a limited number of subgroups, and thus affords the possibility of developing possible applications using far smaller databases. Another noteworthy result lies in the fraction of the core (the centre section E, H, G), which grows as a function of the size of the database: the observed increase is expected to converge asymptotically to 1. Hence, the [G] and [H] circles are absorbed into [E], but the process is symmetrical, in keeping with the proposed universals.

The regularity we observed with respect to types and tokens is also expressed in the relationship between them. Figure 4(b) presents the type/token ratio (TTR) for the four corpora examined. For a given database size, the TTR is similar in all translated corpora and is expected to converge to zero asymptotically for large databases, but this ratio remains higher (by around 20%) for original English, independent of corpus size, pointing to greater linguistic richness.

The Venn diagram then appears to be helpful in pointing to differences between languages, by providing a statistical means of enhancing our fundamental understanding of language(s), both in isolation and in

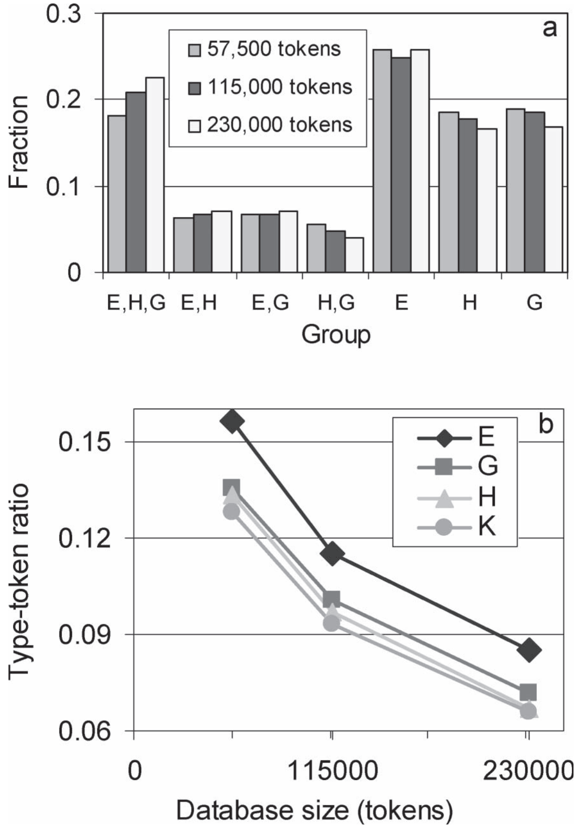

Fig. 4. Universal behaviours as independent of database size. (a) The number of types in each group, displayed as a fraction of the total number of types, for corpora of different sizes. (b) The type/token ratio (TTR) of each corpus vs database size. The TTR is expected to be somewhat higher for original language than for a translation, and to decay to zero as database size increases.

contact. As a statistical measure of relationships between databases, it may be used to distinguish an original corpus from a translated one.

Further research is required to test the potential for generalizing the Venn diagram to higher dimensions. Will the relations hold, for instance, for English and several translated languages? If so, what changes will occur in the scaling of the core? Another possibility is to analyze the dependence of the seven sections as a function of the size of the databases and write a set of rate equations for the likelihood of assigning a new word to a particular section. There are numerous possible directions for future studies. The interface of statistical physics, corpus linguistics and Translation Studies opens a new field for analysis, one whose applications transcend the three disciplines.

REFERENCES

Baker, M. (1993). Corpus linguistics and translation studies – Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology. In honour of John Sinclair* (pp. 233–250). Amsterdam: John Benjamins.

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds), *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* (pp. 17–35). Tübingen: Narr.

Blum-Kulka, S., & Levenston, E. A. (1983). Universals of lexical simplification. In C. Færch & G. Kasper (Eds), *Strategies in interlanguage communication* (pp. 119–39). London, New York: Longman.

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjær & D. Gile (Eds), *Claims, changes and challenges in translation studies* (pp. 1–13). Amsterdam, Philadelphia: John Benjamins.

Choi, S.-W. (2000). Some statistical properties and Zipf's Law in Korean text corpus. *Journal of Quantitative Linguistics*, *7*, 19–30.

Gell-Mann, M. (1994). *The Quark and the Jaguar: Adventures in the Simple and the Complex*. New York: Henry Holt and Company.

Halibard, M., & Kanter, I. (1998). Markov processes and linguistics. *Physica A*, *249*, 525–535.

Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies and Zipf's Law in the Greek language. *Journal of Quantitative Linguistics*, *8*, 175–185.

Jantunen, J. H. (2004). Untypical patterns in translations: Issues on corpus methodology and synonymity. In A. Mauranen & P. Kujamäki (Eds), *Translation universals: Do they exist?* (pp. 101–126). Amsterdam, Philadelphia: John Benjamins.

Kanter, I., & Kessler, D. A. (1995). Markov processes: Linguistics and Zipf's Law. *Physical Review Letters*, *74*, 4559–4562.

Kenny, D. (1998). Creatures of habit? What translators usually do with words. *Meta*, *43*, 515–523.

Laviosa-Braithwaite, S. (1997). Investigating simplication in an English comparable corpus of newspaper articles. In K. Klaudy & J. Kohn (Eds), *Transferre necesse est. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting* (pp. 531–540). Budapest: Scholastic.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L. Havlin, S., Peng, C.-K., Simons, M., & Stanley, H.E. (1995). Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev.* E, *52*, 2939.

Nilsson, P.-O. (2002). Investigating characteristic lexical distributions and grammatical patterning in Swedish texts translated from English. In A. Wilson, T. McEnery & P. Rayson (Eds), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World* (pp. 99–107). Munich: Lincom-Europa.

Tirkkonen-Condit, S. (2004). Unique items – Over- or under-represented in translated language? In A. Mauranen & P. Kujamäki (Eds), *Translation Universals: Do They Exist?* (pp. 177 – 184). Amsterdam, Philadelphia: John Benjamins.

Vanderauwera, R. (1985). *Dutch Novels Translated into English: The Transformation of a Minority Literature*. Amsterdam: Rodopi.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Reading: Addison-Wesley Press.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.