Meta
Journal des traducteurs
Translators' Journal

META

Assessing the Status of Technical Documents as Textual Materials for Translation Training in Terms of Technical Terms

Kageura Kyo

Volume 63, Number 3, December 2018

Traductologie de corpus : 20 ans après

URI: https://id.erudit.org/iderudit/1060172ar
DOI: https://doi.org/10.7202/1060172ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Kyo, K. (2018). Assessing the Status of Technical Documents as Textual Materials for Translation Training in Terms of Technical Terms. *Meta*, *63*(3), 766–785. https://doi.org/10.7202/1060172ar

Article abstract

In this paper, we examine how methods for evaluating corpora in terms of technical terms can be used for characterising technical documents used as textual materials in translation training in a translation education setup. Technical documents are one of the standard types of textual materials used in translation training courses, and choosing suitable materials for learners is an important issue. In technical documents, technical terms play an essential role. Assessing how terms are used in these documents, therefore, would help translation teachers to choose relevant documents as training materials. As corpus-characterisation methods, we used self-referring measurement of the occurrence of terminology and measurement of the characteristic semantic scale of terms. To examine the practical applicability of these methods to assessing technical documents, we prepared a total of 12 short English texts from the six domains of law, medicine, politics, physics, technology and philosophy (two texts were chosen from each domain), whose lengths ranged from 300 to 1,150 words. We manually extracted terms from each text, and using those terms, we evaluated the nature and status of the textual materials. The analysis shows that even for short texts, the corpus-characterisation methods we provide useful insights into assessing textual materials.

Assessing the Status of Technical Documents as Textual Materials for Translation Training in Terms of Technical Terms

KAGEURA KYO
*Graduate School of Education, The University of Tokyo**
kyo@p.u-tokyo.ac.jp

RÉSUMÉ

Dans cet article, nous examinons les méthodes d'évaluation des corpus spécialisés fondées sur l'analyse des termes en vue de la caractérisation des documents techniques destinés à une utilisation comme matériau pédagogique dans l'enseignement de la traduction. Les documents techniques sont le type de matériau textuel le plus utilisé dans les cours de traduction et le choix du matériel approprié pour les apprenants est d'un intérêt majeur. Dans les documents techniques, les termes techniques jouent un rôle essentiel. L'évaluation de la façon dont les termes sont utilisés dans ces documents aiderait les enseignants de traduction à choisir les documents appropriés pour soutenir la formation en traduction. La méthode de caractérisation des corpus que nous proposons combine deux techniques : l'évaluation auto-référentielle de l'apparition des termes dans le corpus et la mesure de l'échelle sémantique qui caractérise ces termes. Pour examiner l'applicabilité pratique de ces techniques pour l'évaluation des documents spécialisés en termes de difficultés pour la traduction, nous avons sélectionné 12 textes anglais courts dans six domaines différents : le droit, la médecine, la politique, la physique, les technologies et la philosophie. Chaque domaine est représenté par deux textes dont les longueurs variaient de 300 à 1150 mots. Nous avons ensuite procédé à l'extraction manuelle des termes depuis le corpus et, enfin, en nous fondant sur l'analyse de ces termes, nous avons évalué la nature et le statut des documents textuels composant le corpus. L'analyse montre que, même pour les textes courts, les techniques de caractérisation du corpus que nous avons adoptées offrent des informations utiles pour l'évaluation des difficultés traductionnelles que peuvent présenter les documents spécialisés lors de leur utilisation comme matériau pédagogique en classe.

ABSTRACT

In this paper, we examine how methods for evaluating corpora in terms of technical terms can be used for characterising technical documents used as textual materials in translation training in a translation education setup. Technical documents are one of the standard types of textual materials used in translation training courses, and choosing suitable materials for learners is an important issue. In technical documents, technical terms play an essential role. Assessing how terms are used in these documents, therefore, would help translation teachers to choose relevant documents as training materials. As corpus-characterisation methods, we used self-referring measurement of the occurrence of terminology and measurement of the characteristic semantic scale of terms. To examine the practical applicability of these methods to assessing technical documents, we prepared a total of 12 short English texts from the six domains of law, medicine, politics, physics, technology and philosophy (two texts were chosen from each domain), whose lengths ranged from 300 to 1,150 words. We manually extracted terms from each text, and using those terms, we evaluated the nature and status of the textual materials. The analysis shows that even for short texts, the corpus-characterisation methods we provide useful insights into assessing textual materials.

RESUMEN

En este trabajo se examina cómo se pueden utilizar los métodos de evaluación de los corpus especializados basándose en el análisis de los términos técnicos para caracterizar los documentos técnicos se que se emplean como materiales textuales en la formación de traductores. Los documentos técnicos son uno de los materiales textuales más utilizados en los cursos de formación en traducción y la selección de materiales adecuados es un tema importante. En los documentos técnicos, los términos técnicos desempeñan un papel esencial. Por lo tanto, evaluar cómo se utilizan los términos en estos documentos ayudaría al profesorado en traducción a elegir los documentos pertinentes como material de aprendizaje. Como métodos de caracterización de corpus, utilizamos la evaluación autoreferencia de la aparición de los términos y la medida de la escala semántica que caracteriza los términos. Para examinar la aplicabilidad de estos métodos a la evaluación de documentos técnicos, se seleccionaron 12 textos cortos en inglés pertenecientes a seis ámbitos, el del derecho, medicina, política, física, tecnología y filosofía (se eligieron dos textos para cada dominio), textos cuya extensión variaba entre 300 y 1150 palabras. Hemos extraído manualmente los términos de cada texto y, basándonos en el análisis de dichos términos, evaluamos la naturaleza y el estatuto de los materiales textuales. El análisis demuestra que, incluso para textos cortos, los métodos de caracterización de corpus que hemos adoptado dan una información útil para evaluar las dificultades traductológicas que puedan presentar los textos especializados utilizados como soportes textuales didácticos.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

documents spécialisés, formation en traduction, couverture conceptuelle, échelle sémantique, mesures quantitatives
specialised documents, translation training, conceptual coverage, semantic scale, quantitative measures
documentos especializados, formación de traductores, alcances conceptuales, escala semántica, medidas cuantitativas

1. Introduction

Specialised documents are one of the standard textual materials used in translation training within a translation education setup. This reflects the fact that there is a strong demand for specialised translation, which "covers the specialist subject fields falling under non-literary translation, the best known of which include science and technology, economics, marketing, law, politics, medicine and mass media" (Gotti and Šarčević 2006: 9).

One of the most important issues in translating specialised documents is handling technical or domain-specific terms (henceforth simply "terms") properly, which ranges from translating terms properly and consistently to managing terminology resources for translation projects (Bowker 2015; Cabré 2010). Understanding how to handle terminology in translation is an essential part of domain competence, one of the six competences required for professional translators listed in ISO 17100.[1]

In a translation training setup, choosing suitable materials for learners is an essential issue. In the case of choosing specialised documents as materials for translation training, how terms are used in the documents constitutes an essential factor, as terms critically determine, or reflect, the nature of specialised documents, and, as just stated above, translators should learn how to handle terms. Assessing how terms

are used in specialised documents would help translation teachers to choose relevant documents as training materials.

The relationship between terms and texts has been explored so far in the fields of terminology, corpus linguistics, quantitative linguistics and natural language processing. The knowledge accumulated in these studies can provide useful insights into how terms and texts are related and offer methods for characterising specialised documents from the point of view of the terms they contain; they can provide useful information for examining the relevance of the documents as textual materials for translation training. However, not much work has been carried out so far that proposes that this knowledge can possibly be used for choosing specialised documents as textual materials for translation training.

Against this backdrop, this paper examines and evaluates how methods for analysing corpora can be applied to characterising specialised documents used as textual materials for translation training in a translation education setup. Note that the goal in this paper is to propose methods of characterising individual technical documents in terms of technical terms and to validate the practical feasibility of these methods.

The rest of the paper is organised as follows. In Section 2, we will briefly look at the relationship between terms and texts in the context of translation and summarise the desiderata for evaluating specialised documents, especially from the point of view of occurrence of terms. We will also review technical studies on the relationship between terms and texts, through which we select two measures that can be used for evaluating the characteristic of specialised documents in relation to terms. Section 3 elaborates on these two measures. In Section 4, we apply these measures to evaluating a range of short documents that cover different specialised topics and we also examine the feasibility of adopting these methods. Section 5 concludes the paper.

2. The status of terms in relation to texts

Here we first clarify the desiderata for evaluating specialised documents with respect to terms, by examining the status of terms in specialised documents. We will then briefly observe existing work and identify useful approaches. To maintain wider applicability, we focus on the methods of evaluating textual materials that rely only on a given document, rather than assuming the use of external resources. It is also desirable that the methods do not require intensive human labour or special skills. For these reasons, we focus on quantitative textual analysis as the potential methods of choice.

2.1. Terms in specialised translation

Handling technical terms is one of the most important elements in specialised translation. Although what characterises specialised translation is not limited to terms (Byrne 2006: 3-4), terms have been a focal topic in studies of specialised translation (Cabré 2010; Montero Martinez and Faber 2009; Rogers 2008). From the point of view of the study of terminology, translation is widely understood as one of the most important areas of application (Bowker 2015; Sager 1990). Though existing studies that deal with terms in specialised translations cover a wide range of topics, such as

translating terms, terminology management, teaching terminology in the context of translation, and automatic terminology processing, from the point of view of our objective in this study, it is sufficient here to confirm two essential roles of terms and terminology in specialised documents.

First, terms represent concepts inside a domain (Felber 1984; Sager 1990). A concept in a domain has status within the totality of the conceptual system of the domain, and can therefore be fully understood in relation to neighbouring concepts. This conceptual system is represented by the terminology of the domain (Kageura 2012). Although a specialised document in a domain only refers to a part of this conceptual system and thus uses a limited number of terms of the domain, it is not sufficient to know only these terms and concepts to handle the document. Understanding specialised documents requires knowledge of the overall conceptual system and thus of the terminological structure of the domain. For instance, take an article in the field of natural language processing that deals with morphological analysis. The article may not contain the term (and therefore the concept of) "parsing," but knowing the relationship between morphological analysis and parsing is a *sine qua non* for properly understanding the content of the article. In specialised translation, which requires an understanding of the source material, knowing the target language equivalents of technical terms is not sufficient. Translators should also be knowledgeable about the group of related concepts and terms that do not occur in the document (Cabré 2010; Wright and Wright 1997).

Second, in specialised documents, terms constitute an essential pillar upon which arguments or textual discourse are constructed (Sager, Dungworth, *et al.* 1980). In some cases, though it is not so common, the translation of terms may be affected by the nature of the document, including the structure of discourse. After all, although terms tend towards the rigidity of their designations, they also have a certain degree of flexibility, reflecting the fact that terms share forms with non-terms (Rey 1995; Temmerman 2000). Specialised translators should therefore be knowledgeable about how terms are used within the discourse or how they are used to develop arguments or the structure of discourse in texts.

These two points, which are widely accepted and do not seem necessary to restate explicitly, give important insights into what is desirable to take into account when we evaluate specialised documents as training materials from the point of view of terms. First, it may not be sufficient to observe and evaluate terms that actually occur in texts, as the understanding of these terms depends to some extent on related terms that may not occur in the texts. Second, taking simple occurrence statistics may not suffice, because how terms are used within the discoursal structure of the document affects the characteristics of the documents. For the first issue, it may be useful to use external resources such as the terminology of a domain or lexical resources that contain such information as familiarity or specificity. Unfortunately, such relevant resources are often unavailable. For the second issue, detailed manual analysis can provide a solution, but it is time-consuming and requires skills and experience.

We instead propose a more practical solution, namely to introduce summary measures that can capture (a) the status of terms occurring in a document in relation to a terminological system that is expected to exist behind the scenes, and (b) the characteristics of term usage within the discoursal structure, which can be calculated using a given document, without resorting to external resources or intensive manual

work. Before proposing concrete measures, we shall briefly examine some related work.

2.2. Related work on the relationship between terms and texts

In the field of Translation Studies, Nord (1988/1991) comprehensively examined issues related to text analysis in translation, in which requirements for choosing textual materials for translation classes were also discussed. The level of requirements given in Nord (1988/1991), though essential, does not match the immediate objective of our study; the results of our study, if useful, should be used within the framework proposed by Nord (1988/1991). Work on characterising texts in translation studies has mostly targeted translated texts rather than textual materials for translation training, and quantitative approaches are no exception to this tendency (Oakes and Ji 2012). It should also be noted that the quantitative text analysis methods used in translation studies are mostly standard ones proposed in quantitative linguistics in general.

In some other application areas such readability studies, characterising texts is a critically important topic. Several well-known general measures, such as the Flesch-Kincaid measure, have been proposed and used so far.[2] These general measures, however, do not directly address the issue of specialised documents and terms. Stylometry and studies of authorship attribution are also concerned with characterising texts. They have proposed a variety of features, including character, lexical, syntactic, semantic and other features for evaluating textual characteristics (Stamatatos 2009; Yule 1944). Lexical features are relevant to our study, but they can be more generally understood within the framework of lexical statistics rather than in relation to the specific application of stylistic analysis or authorship attribution. Also, as the task of authorship attribution often aims to cluster documents without directly characterising individual documents, the methods proposed in authorship attribution research are generally not of direct use for our aim. Some work addresses the issue of technical specificity or terminology load in the context of characterising specialised documents (Jones 1995; Asaishi and Kageura 2016). There is also work on descriptive studies that examine terms in context (Pearson 1998). These studies, although interesting, are specific to their own research context and do not necessarily provide generic measures or methods that fit our aim, or the measures used in these studies would be better understood within a more generic context.

The relationships between terms and corpora or textual data have so far been explored from several different points of view in the field of computational terminology. Extracting monolingual or multilingual terms from textual corpora is one of the most active application-oriented research areas, within which a variety of methods have been proposed (Kageura and Umino 1996; Ahmad and Rogers 2001; Heylen and de Hertog 2015). Another potentially related research area is information retrieval (IR), in which terms or keywords are studied in order to facilitate retrieving, classifying or clustering documents (Manning, Raghavan, *et al*. 2009). As the relationships between terms and documents have duality (Aizawa 2000), one can both use texts to characterise terms and use terms to characterise texts. Nevertheless, there is not much work that explicitly aims at characterising the nature of texts by means of terms in the context of computational terminology and IR. This is probably due

to the fact that the focus is on terms rather than texts in the case of computational terminology and on the application performance in the case of IR.

In the field of quantitative and computational linguistics, we can identify several studies that will be useful to our objective. As we are concerned with terms and texts, a methodologically relevant area is quantitative analyses of words in texts or lexical statistics. A variety of perspectives has been proposed and a wide range of methods to analyse different aspects of distribution of words in texts has been accumulated in this research area.

One way of characterising texts by means of terms is to observe the basic quantitative nature of terms. The following quantitative measures are widely used for basic characterisation of texts by means of the distribution of words (Baayen 2001; Gries 2009):

- The ratio of tokens: this quantity is used with some classifying features, such as POS categories (for example, the ratio of nouns in a running text).
- The ratio of types: the ratio of word types as counted by the number of different words that belong to a certain group, to all word types.
- Type-token ratio: the number of word types divided by the number of word tokens. The reciprocal of this measure is the mean frequency per word.
- Type-frequency distribution: the distribution that captures the relationship between the frequency of occurrence and the number of word types that occur with a given frequency. The parameter of parametric distributions – such as the so-called Zipfian distributions (Zipf 1935) or Poisson distributions – fitted to empirical data is sometimes used as a measure to characterise the nature of the distribution of the words and thus the text (Popescu 2009).

These basic measures, though generally held essential and widely used as a first step towards analysing texts and words quantitatively in general, do not satisfy the desiderata we consolidated in Section 2.1, as they only capture quantities of elements that actually occur in a given text, and also, they cannot deal with the discoursal aspect of the occurrence of words. Such measures as variance (of frequencies or intervals) can also be observed, but they are in general observed to be unstable and tend to change due to a small number of outliers (such as a single very frequent word or a single occurrence that is far away from other occurrences).

Approaches that satisfy our desiderata can be found within more technically-loaded studies on word distribution. First, we can see that the so-called zero-frequency problems addressed in quantitative and computational linguistics address an issue related to evaluating the status of terms occurring in a document in relation to the terminological system. It has been recognised that if we extend the size of textual data or a corpus, we tend to observe new words that have not occurred in existing data (Baayen 2001; Manning and Schütze 1999). From the point of view of constructing probabilistic language models, this issue leads to the reassignment of smaller probabilities to existing words, as this is necessary to reserve a certain amount of probability mass to unseen words that may occur when the corpus size is extended. This is directly related to the problem of estimating how many unseen words there potentially are, the problem addressed by some quantitative linguistic studies (Baayen 2001). If we can estimate the potential size of terminology from the distribution of terms in a given document, it is possible to evaluate the status of the set of terms used in the document. The potential size of the terminology, which is used to evaluate the

status of terms actually occurring in a document, is in fact estimated from the document. It is thus called a self-referring measurement. Kageura and Kikui (2006) applied this idea to evaluating the status of a travel corpus. Though their evaluation is based on general words and not on technical terms, the basic idea can be applied, at least theoretically, to evaluate the status of the set of terms occurring in a given text with respect to potential terminology and also to characterise the specialised document in view of the status of terms occurring in it. We will elaborate on the technical aspects of this approach in Section 3.1.

Second, some recent studies in quantitative linguistics, especially those that make use of the methodologies and ideas developed in statistical physics, have analysed occurrence patterns of keywords by looking at word sequences in a text as time series data (Ortuño, Carpena, *et al*. 2002; Zhou and Slator 2003; Herrera and Pury 2008; Mehri and Darroneh 2011; Yang, Lei, *et al*. 2013). Within this research trend, Montemurro and Zanette (2010) proposed a characteristic semantic scale of words within the discoursal structure of a text. The basic idea is to experimentally determine a textual span that maximises the information burden of a word. This can be interpreted as reflecting the characteristics of term usage within the discoursal structure of the text, and can be applied to evaluating how terms relate to the discoursal structure, without resorting to qualitative analysis. We will detail this approach in Section 3.2.

3. Measures for evaluating the status of textual materials

We have consolidated the desiderata for measures characterising specialised texts with respect to terms and we have identified quantitative approaches that can be useful as such measures. Here we elaborate on these measures more formally and also on their interpretations.

3.1. Coverage of conceptual range by terms in documents

In Section 2.1, we stated that it is necessary for a translator to be knowledgeable about concepts represented by terms that do not occur in a given specialised document, but are related to those that do. In quantitative terms, this statement can be mapped to more simplified questions as follows:

- What is the expected size of the terminology to which the terms that occur in the document belong?
- What is the ratio of the terms occurring in the document to the expected size of the terminology?

For instance, if the expected size of the relevant terminology is small and the ratio of occurring terms among this terminology is high, then relatively less knowledge of the background conceptual system is required to translate the document. Alternatively, if the expected size of relevant terminology is large and the ratio of occurring terms is low, then translators should know a wider range of concepts to translate the document. Though this interpretation does not take into account qualitative factors such as complexity or familiarity of concepts, etc., it still holds as a rough approximation for understanding the status of terms that occur in a given document with respect to the conceptual system of the specialised knowledge, a part of which is handled in the document.

In terms of quantitative analysis, we can adopt the following approach to answer these questions:

- to estimate how many terms would occur when the size of the given document is assumed to be extrapolated to infinity; and
- to evaluate the ratio of terms that actually occur in the document to the estimated size of terminology. We can call this ratio the coverage of conceptual range by terms.

The core technical issue is related to the first point, that is, to estimate the population size of terminology based on the terms in the data.

A number of methods for estimating the population size have been proposed so far (Efron and Thisted 1976; Tuldava 1995; Baayen 2001). We adopt here the Large-Number-of-Rare-Events (LNRE) estimation method (Khmaladze 1987[3]; Baayen 2001). The method has been successfully applied to characterise corpora and terminologies (Kageura and Kikui 2006; Kageura 2012; Miyata and Kageura 2016). It is also convenient from the application point of view, since a statistical package called zipfR (Evert and Baroni 2007) is available on the general data analysis environment *R*.[4]

Informally speaking, the method estimates the population size of vocabulary under the assumptions that words are randomly distributed in a corpus and that the distribution follows a certain form, which can be captured by certain parametric distributions. The basic framework can be explained more formally as follows (Baayen 2001). First, let us define the set of notations as follows:

S	:	the population number for the terms (to be estimated by using a given text);
t_i	:	the i-**th** term type (i = 1, 2, … S);
p_i	:	the population probability of a term type $\mathbf{t}_i$;
N	:	the length of the text as counted by the number of term tokens;
m	:	the frequency of occurrence for a term;
$\mathbf{V}(m, N)$	:	the number of term types occurring m times in the text of length N;
$\mathbf{V}(N)$	:	the number of term types occurring in the text of length N.

Under the randomness assumption, the expected number of term types that occur m times in a given text of size N is given by:

$$\mathbf{E}[\mathrm{V}(m, N)] = \sum_{i=1}^{S} p_i^m (1 - p_i)^{N-m}.$$

The number of term types that occur in a given text of size N is given by summing up $\mathbf{E}[\mathrm{V}(m, N)]$ for all t_i and $m > 0$, which is equivalent to subtracting the number of term types that do not occur in the text from the population size S:

$$\mathbf{E}[\mathrm{V}(N)] = \sum_{m=1}^{N} \sum_{i=1}^{S} p_i^m (1 - p_i)^{N-m} = S - \sum_{i=1}^{S} (1 - p_i)^{N-m}.$$

Although we will not get into a detailed technical discussion here, this framework suggests that we will observe new words roughly in accordance with the ratio of hapax legomena, or words that occur only once in the existing data, when the size of the text is extended by one-word token. Also, when N is extrapolated to infinity, $\mathbf{E}[\mathrm{V}(N)]$ gives the population number of types S. So the necessary task is to estimate S, based on a given distribution of terms.

Note that these formulae assume that we know the population number of terms. In reality, this is exactly what we want to estimate, and what is given is the number of term types that actually occur in a given text and their frequency distributions. To estimate the population number of terms from what is observed in the data by tracing the reverse path from this, it is assumed that the distribution of terms follows such parametric distributions as Zipfian distributions, Log-normal distributions or inverse Gauss-Poisson distributions, all of which capture highly skewed distributions typically observed in word frequency distributions. As we will see below, in the experiment, we focus on the LNRE method with a Zipf-Mandelbrot distribution (Mandelbrot 1953; Evert 2004). A Zipf-Mandelbrot distribution is defined as follows:

$$\mathbf{p}(t_i) = \frac{C}{(\mathbf{r}(t_i) + b)^a}$$

where $\mathbf{p}(t_i)$ indicates the ratio or probability of the term t_i, $\mathbf{r}(t_i)$ indicates the rank of the term t_i (with the most frequently occurring term being rank 1), and a, b, and C are parameters. Evert (2004) details the Zipf-Mandelbrot LNRE method. The randomness assumption discards the local and discoursal dependency of the occurrences of terms, but what is estimated by the LNRE methods is the conceptual range of terms in the document as a set, rather than the pattern of occurrences. For that, the randomness assumption adopted in the LNRE methods is not unreasonable.

S gives the number of term types necessary to cover the topic dealt with in a given text and approximates the range of concepts when the topic is fully explored. Observing what percentage of **S** is covered by the terms actually occurring in the text gives the status of texts in relation to the range of the concepts. As S is estimated for a given text by using the actual distributional paper, and once we obtain S, the status of the text is evaluated in turn. We refer to this method as a self-rereferring measurement for the occurrence of terminology.

3.2. Semantic scale of terms in a discoursal structure of documents

In Section 2.1, we confirmed that it is desirable for us to be able to evaluate how terms are related to the discoursal structure of a given document. For instance, a term that represents a very specific concept may occur in a very limited textual area, while another term may be used throughout the document. If most of the terms are used throughout the document, the document may well address a single complex topic the whole way through, while if only a small portion of terms are used throughout the document and other terms occur only locally in different locations, we can guess that the argument in the document consists of several different branches or segments. If we can measure over what range of text a concept represented by a term stretches, we will be able to grasp the characteristics of discoursal structure in a given specialised document.

Observing the occurrence interval is a clue that immediately comes to mind, but to define a relevant semantic scale in discourse, directly correlated with the concentration of the occurrence of terms by means of interval information is not as easy as it first appears (Ortuño, Carpena, *et al.* 2002; Zhou and Slator 2003; Herrera and Purry 2008; Mehri and Darroneh 2011; Yang, Lei, *et al.* 2013). Naively, for instance, it is not easy to compare a term occurring three times in a text with intervals of 5

and 10 for another term occurring twice with an interval of 8. It is necessary to introduce a measure that enables the comparison of terms occurring with different frequencies from the point of view of semantic scale.

The measure proposed by Montemurro and Zanette (2010), originally intended for weighting keywords in a document, can be used for this aim. Their basic idea is to evaluate the amount of information for each term, not directly, but in comparison with a randomised version of the text. As the randomised version of the text contains the same words with the same frequencies as the original text, it is understood that the information, measured as the difference between these two texts comes from the discoursal factors of the text. Note that this information itself does not directly give a term's semantic scale. The second idea is to change the unit of randomisation and evaluate the difference between the information of a word in the original text and the randomised text for each method of randomisation. The unit of randomisation that maximises the information of the term is then chosen as the semantic scale of the term within the discoursal structure of the document.

Formally, the semantic scale of a term can be calculated as follows. First, we define the notation as follows:

$\boldsymbol{D}$	:	the number of discoursal units we identify;
$\boldsymbol{f}$	:	the frequency of a term $\boldsymbol{t}_i$ in the text;
$\boldsymbol{f}_j$	:	the frequency of a term $\boldsymbol{t}_i$ in the ***j*-th** textual unit;
$\boldsymbol{N}$	:	the length of the text as counted by the number of tokens of all the terms;
$\boldsymbol{J}$	:	the random variable for the textual unit, taking 1, …, $\boldsymbol{D}$.

Note that, if we set the discoursal unit as a paragraph, $\boldsymbol{D}$ equals the number of paragraphs in the text. If we set the unit as a sentence, $\boldsymbol{D}$ equals the number of sentences.

The probability that a term is $\boldsymbol{t}_i$, $\mathbf{p}(\boldsymbol{t}_i)$, is given by the relative frequency of $\boldsymbol{t}_i$ among all the term tokens, $\boldsymbol{f}/\boldsymbol{N}$. The probability of observing the ***j*-th** unit given a token of arbitrary terms, namely $\mathbf{p}(\boldsymbol{j})$, is represented by the relative length of the ***j*-th** unit. The probability that the unit in which a token of $\boldsymbol{t}_i$ occurs is ***j*-th**, namely $\mathbf{p}(\boldsymbol{j}|\boldsymbol{t}_i)$, is given by $\boldsymbol{f}_j/\boldsymbol{f}$. The semantic scale for a term $\boldsymbol{t}_i$ is calculated as follows:

a) Divide a given text into $\boldsymbol{D}$ units;
b) For each division of units,
   – calculate the mutual information, $\mathbf{MI}(\boldsymbol{t}_i,\boldsymbol{J})$, between $\boldsymbol{t}_i$ and $\boldsymbol{J}$ for the original text:

$$\begin{aligned}\mathrm{MI}(t_i, J) &= \mathrm{p}(t_i)\sum_{j=1}^{D}\mathrm{p}(j|t_i)\log_2\frac{\mathrm{p}(t_i)\mathrm{p}(j|t_i)}{\mathrm{p}(t_i)\mathrm{p}(j)}\\ &= \mathrm{p}(t_i)\sum_{j=1}^{D}\mathrm{p}(j|t_i)\log_2\frac{\mathrm{p}(j|t_i)}{\mathrm{p}(j)}\end{aligned}$$

   – calculate the mutual information between $\boldsymbol{t}_i$ and $\boldsymbol{J}$ for a randomly reordered text,[5]
   – take the difference **ΔMI** between the mutual information for the original text and for the randomised text, and;
c) take the unit that maximises **ΔMI** as the semantic scale of $\boldsymbol{t}_i$ for the text.

As can easily be seen from the above procedure, this measure cannot be defined for hapax terms, because occurrence intervals cannot be defined for hapax terms and thus the randomised texts and the original texts carry the same information. It could also be unreliable for low-frequency terms, as chance factors can affect their intervals.

Despite these technical shortcomings, we contend that it is important to examine its applicability as it theoretically fits the characterisation of terms in the discoursal structure of documents.

4. Data and experiments

In the previous two sections, we have established the desiderata for evaluating specialised documents as translation materials and elaborated on the quantitative measures that can capture, but only approximations, the desirable features to take into account when evaluating the status of specialised documents in relation to terms within a translation training setup. This does not guarantee that these measures can be used on textual materials used for translation training. The following two points need to be validated:

- Textual materials used for translation training are typically shorter than most texts to which these methods have been applied. It is therefore necessary to check the applicability of these methods to shorter texts.
- Even though we argued for the relevance of these measures in evaluating specialised documents as training materials, whether the materials chosen by using these measures would actually help training performance needs to be empirically validated.

We will examine the first issue here. As the second issue involves a qualitatively different phase of research, we can safely leave it to future research.

4.1. Data

For data, we chose 12 short texts that belong to six different domains, namely law (two excerpts from a research monograph, which we will call L1 and L2 for brevity), medicine (two abstracts from an academic journal; M1 and M2), politics (two articles taken from an online journal; P1 and P2), physics (two abstracts from an academic journal; S1 and S2), technology (two articles from an online journal; T1 and T2), and philosophy (two entries from Wikipedia; W1 and W2).[6] Their lengths range from 300 to 1150 words. They are taken from a repository of texts for a translation course taught by the author. For the same domain, we chose texts that belong to the same register and have approximately the same length. The medicine and physics texts are highly specialised, while those for law and philosophy are less specialised in comparison, and those for politics and technology are more journalistic and even less specialised.

Note that our task here is to validate the applicability of the measures to short texts, and not to characterise the texts. We use these differences in register and in the degree of specialisation as guidance for checking the nature of the measures. This is not completely justifiable logically because the correlation between the characteristics of texts with respect to terms and characteristics of texts identified from other points of view is not guaranteed. Also, if we recall our original objective, term-based measures should be discriminative of texts that belong to the same register. Nevertheless, at this stage, we believe that using texts with a different degree of specialisation and from a different specialised domain would help to further understand the nature of the measures we introduced.

The basic statistics are given in Table 1. We can observe that the text pairs in the same domains have similar characteristics for other quantities, such as the number of word types, the mean frequency per word, the number of sentences, and the mean length of a sentence, though we can observe larger differences in some domains than in others.

TABLE 1

Basic quantities of the 12 texts

text	WTKN	WTYP	WTKN/WTYP	SNT	MLSNT
L1	719	314	2.29	16	44.9
L2	642	264	2.43	14	45.8
M1	385	205	1.88	27	14.3
M2	374	193	1.94	27	13.9
P1	1127	426	2.65	43	26.2
P2	1144	415	2.76	53	21.6
S1	342	189	1.81	13	26.3
S2	306	170	1.80	12	25.5
T1	668	292	2.29	20	33.4
T2	635	298	2.13	27	23.5
W1	616	229	2.69	18	34.2
W2	552	239	2.10	18	30.7

WTKN stands for the number of word tokens or the length of the text as counted by the number of running words; **WTYP** stands for the number of word types; **WTKN/WTYP** shows the mean frequency per word; **SNT** stands for the number of sentences; **MLSNT** stands for the mean length of a sentence.

4.2. Basic quantities of terms in the data

From these texts, we manually identified the technical terms. We defined a technical term here as a simple or complex lexical or phrasal unit that represents a unified concept and that needs to be consistently translated into a specific target language (TL) expression. We included proper names in the scope of technical terms. Terms were identified by the author and a teaching assistant who helped in a translation class taught by the author. Final decisions were made by the author. Given that automatic term extraction methods have generally achieved practically usable performance (Heylen and de Hertog 2015), we could apply automatic term extraction instead of manual extraction to identify terms in texts. Here, we resorted to manual extraction as we already had extracted terms before we started this experiment.

The basic statistics for the terms in the 12 texts are given in Table 2.

Table 2
Basic statistics of terms in the 12 texts

Text	N	V	N/V	NNORM	VNORM	TRTKN	TRTYP
L1	115	66	1.74	160	63	0.16	0.21
L2	118	60	1.97	184	68	0.18	0.23
M1	101	50	2.02	262	73	0.26	0.24
M2	77	49	1.57	206	76	0.21	0.25
P1	79	43	1.84	70	30	0.07	0.10
P2	83	43	1.93	73	31	0.07	0.10
S1	62	42	1.48	181	67	0.18	0.22
S2	75	45	1.67	245	79	0.25	0.26
T1	73	25	2.92	109	26	0.11	0.09
T2	55	26	2.12	87	26	0.09	0.09
W1	105	62	1.69	170	81	0.17	0.27
W2	96	62	1.55	174	78	0.17	0.27

N stands for the number of term tokens and **V** stands for the number of term types, so that **N/V** shows the average frequency of terms. **NNORM** and **VNORM** respectively give the number of term tokens and term types when the length of texts is normalised to 1,000 under a simple linear assumption, for a coarse intuitive comparison. **TRTKN** gives the ratio of term tokens to word tokens (and thus equals **NNORM**/1000), and **TRTYP** gives the ratio of term types to word types. Note that the last two quantities are an approximation, as words are delimited by spaces and punctuation while terms can be complex.

From these basic statistics for the terms, we can observe several characteristics of texts with respect to terms. We can point out the following tendencies, among others:

- Generally speaking, two texts that belong to the same domain show similar tendencies. This means that the usages of terms in specialised documents of the same domain and with the same register are generally similar.
- The ratios of term tokens to word tokens are generally higher for more specialised texts: the highest two are M1 and S2, then followed by M2, L2, and S1.
- The ratios of term types to word types are higher for more specialised texts, though the highest two, W1 and W2, are from Wikipedia.
- The mean frequency of a term shows a weak tendency for less specialised texts to have a higher mean frequency per term; in other words, a term tends to be used repeatedly in less specialised texts. However, this tendency relies heavily on the existence of T1 and T2, and in fact is not clear.

Though these summary statistics prove very useful for identifying the overall general characteristics of texts in relation to terms, they do not provide information that meets the desiderata we postulated in Section 2.1.

4.3. Applicability of the two quantitative measures

We shall first reconfirm here the two desiderata for characterising specialised documents by means of the terms they contain, from the quantitative point of view.

The first issue, namely the status of terms that occur in documents with respect to the underlying conceptual system, can be intuitively related to the distribution of terms as follows. Take, for instance, a text that contains 60 term types, each of which

occurs 3 times, and another text of the same length that contains 60 term types, of which 50 occur only once, 1 occurs 40 times and the remaining 9 occur 10 times each. Both texts have 180 term tokens. So the basic quantities we observed are the same (though in this case, the variances differ greatly). Intuitively, the argument in the first document is more or less self-sufficient with respect to concepts, while that in the latter document seems to imply a potentially wider range of concepts (Kageura 2012). We need to delve into the frequency distribution of terms rather than resorting to the basic summary measures.

The second issue, specifically how terms relate to the discoursal structure of the document, is related to the concentration of occurrences of terms. Take, for instance, a text that contains 60 term types, each of which occurs 3 times a more or less equal intervals in the text, and another text that contains 60 term types, each of which occurs again 3 times, but most terms occur in concentrations. Naturally, how the discoursal structure is organised with respect to the concepts represented by these terms differs greatly between these two texts. Basic summary measures fall short of capturing this aspect of term occurrences. As stated, detailed qualitative analysis can be useful for characterising the discoursal structure, but it is both time consuming and judgement is liable to become inconsistent.

Taking these into mind, let us examine below whether these two measures can be reasonably applied to short texts and, if so, what kind of insights we can gain from these measures.

4.3.1. Coverage of conceptual range

We applied the Zipf-Mandelbrot LNRE method (Baayen 2001; Evert and Baroni 2007) using the *R* zipfR package to the 12 texts. The results are shown in Table 3.

TABLE 3
Results of the application of the Zipf-Mandelbrot LNRE method

Text	S	V	V/S	p
L1	299	63	0.21	0.13
L2	182	68	0.37	0.90
M1	167	73	0.43	0.39
M2	201	76	0.37	0.17
P1	187	30	0.16	0.36
P2	279	31	0.11	0.88
S1	149	67	0.45	0.02
S2	243	79	0.33	0.43
T1	43	26	0.60	0.32
T2	74	26	0.35	0.92
W1	260	81	0.31	0.03
W2	372	78	0.21	0.02

S shows the estimated population number of terms and **V** shows the real number of terms in the texts. **V/S** therefore indicates the ratio of terms occurring in the texts to the terminology population, so it can be interpreted as the coverage of conceptual range. Here, **p** stands for the p-value, which shows the validity of the application of the method (the higher the better in this context).

Let us first evaluate the technical applicability of the method to the short data. The p-values vary greatly depending on the text, ranging from 0.02 (the model may not be so valid) to 0.92 (the model fits extremely well). As word frequency distributions are characterised by a large number of items that occur with a very low frequency, the p-values generally tend to be very small in many applications of the LNRE methods (cf. Baayen 2001; Kageura 2012). Compared to existing cases in which the LNRE methods were applied, the p-values for the 12 texts are not bad – indeed, they are very good. Although we have to be careful with the treatment of p-values as they tend to become higher for smaller data, we can conclude that the LNRE methods can be applied to short texts to the same extent that they can be applied to large textual corpora.

Turning to **V/S**, the coverage of the conceptual range, we can observe some interesting new characteristics of the texts. First, we can still identify affinities between two texts in the same domain, although there seems to be a wider range of diversity than we can observe through basic statistics. More importantly, we now understand the following characteristics of the 12 texts:

- Although P1 and P2 contain the smallest number of terms (by type), with the exception of T1 and T2, the population terminology sizes required for understanding P1 and P2 are comparatively larger. Indeed, the terms that do occur represent only 16% and 11% of the background concepts that are taken to be necessary for an adequate comprehension of P1 and P2, respectively.
- Conversely, an acceptable understanding of the science and technology texts (T1, T2, M1, M2, S1, and S2) requires a comparatively smaller number of extratextual terms (**V/S** for L2 is higher than that for T2 and S2, but this is the only exception). It can be understood that these more specialised documents assume a more specific background conceptual system.
- For L1 and W2, and to a lesser extent for P1 and P2, a wider range of concepts necessary for comprehension are not explicitated in these texts, and thus prior knowledge is required. L2 and W1 are closer to more highly specialised texts.

In a real environment within which training materials are chosen, we may well need to characterise and differentiate between specialised documents that belong to the same domain. The current experimental setup is simpler, since we compare texts across different domains with different registers. Still, the tendencies we summarised above indicate that this measure can yield very interesting and important information for choosing specialised documents as textual materials for translation training.

4.3.2. Semantic scale of terms in discourse

To calculate the semantic scale, we need to set the smallest unit based on which larger units are constructed. Montemurro and Zanette (2010) equi-partitioned texts mechanically into an arbitrary number of units. Asaishi (2017) used the paragraph as the basic unit in analysing the semantic scale of high-school science textbooks. As the texts we use are short and dense, we adopt the sentence as the basic unit for observing the semantic scale; thus, the semantic scale is given by the number of successive sentences. As the semantic scale cannot be measured for terms that occur only once, we focus on terms that occur more than once in each text. The results are shown in Table 4.

TABLE 4
Semantic scale of terms in the 12 texts

Text	V2	MN	MX	WMN	WMX	RMN	RMX
L1	115	66	1.74	160	63	0.16	0.21
L2	118	60	1.97	184	68	0.18	0.23
M1	101	50	2.02	262	73	0.26	0.24
M2	77	49	1.57	206	76	0.21	0.25
P1	79	43	1.84	70	30	0.07	0.10
P2	83	43	1.93	73	31	0.07	0.10
S1	62	42	1.48	181	67	0.18	0.22
S2	75	45	1.67	245	79	0.25	0.26
T1	73	25	2.92	109	26	0.11	0.09
T2	55	26	2.12	87	26	0.09	0.09
W1	105	62	1.69	170	81	0.17	0.27
W2	96	62	1.55	174	78	0.17	0.27

V2 shows the number of terms for which a semantic scale was calculated, that is to say the number of terms that occur more than once in the text. **MN** shows the mean semantic scale as counted by the number of sentences. For instance, in M2, the most informative unit of text with respect to technical terms consists on average of 5.5 sentences. **MX** shows the maximum semantic scale. **WMN** and **WMX** show the semantic scale by means of the number of words, by multiplying **MN** and **MX** by **MLSENT** in Table 1, respectively. This is done because the mean sentence lengths differ greatly among the texts, and it was thought useful to give the semantic scales by means of the number of words. **RMN** and **RMX** show the relative scale, **MN/SNT** and **MX/SNT**, respectively (**SNT** is given again in Table 1). They show the semantic scale relative to the length of the text. As L1 and L2 are excerpts from a monograph and cannot be regarded as independent texts, **RMN** and **RMX** may not be meaningful.

From Table 4, we can observe the following points, among others. Note that some of these cannot be observed by simple statistics, namely, the number of term tokens and term types given in Table 2.

- First, the pair of texts for each domain shows a reasonably similar mean semantic scale, both in absolute and relative scales, with the exception of the pair W1 and W2.
- In an absolute scale, P1 and P2 have the largest mean semantic scale: an average meaningful chunk in discourse is around 7 sentences and the maximum semantic scale 17 and 25. Interestingly, these are the two texts with the smallest **RMN** and **RMX**. This may imply that the discoursal unit is not as dense as the other texts and the core thread of the discourse is less clear. Recall that these are taken from online articles on political issues. What can be interpreted through the semantic scale measures intuitively fits our common-sense understanding of these texts relative to the other texts used here.
- **WMN** shows that P1 and P2, T1 and T2, and L1 and L2 have similar scales. For **WMX**, P1 and P2 stand out. Recall that T1 and T2 are also more journalistic articles, and L1 and L2 are from a research monograph.
- M1, M2, S1 and S2, though their similarity is not displayed for **MN** and **MX**, show very similar scales for **WMN**. They also show a clear contrast with the texts in P, L, and T. The figures imply that they have a tight discourse in absolute scale, though the relative mean scale shows that they are about the same as the other documents. This therefore may be a reflection of the fact that they are abstracts of the articles.

If we take a closer look at Table 4, other characteristics may be pointed out. As of now, we contend that the above observations already show that the semantic scales can shed light on the characteristics of texts from the point of view of discoursal structure.

5. Conclusions and outlook

In this paper, we have shown that the methods introduced in quantitative corpus-based terminological processing can be used to reveal hitherto unaddressed aspects of specialised documents in relation to terms, which can provide useful information in the choice of textual materials for translation practice. The measurement of the coverage of the conceptual range shows the range of concepts that should be grasped in order to understand the text. The semantic scale revealed the basic units of discourse constructed by means of technical terms in specialised documents. We have validated that the methods, so far mostly applied to larger texts, can be applied to much shorter texts – typically the types of texts chosen for translation training classes.

Although the two measures, we contended, reflect two important elements of the characteristics of specialised documents in relation to terms, additional factors need to be taken into account for the measures to be technically wholesome. We have already introduced some of these factors in Section 4. For instance, we took into account the average length of sentences for comparing texts the average length of which differ greatly. It is necessary to systematically clarify the factors that need to be normalised in applying these measures.

What we have shown are the theoretical relevance of the measures and their technical applicability to shorter texts. For these measures to prove truly useful for guiding the choice of specialised documents in translation training practice, we need to analyse the relationships between actual translation performance by trainee translators and the nature of specialised documents as characterised by these measures. We have started a preliminary experiment to validate the practical applicability of these measures within a real translation training setup.

ACKNOWLEDGEMENTS

This study is partly supported by JSPS Grant-in-Aid (A) 25240051, "Archiving and using translation knowledge to construct collaborative translation training aid system."

NOTES

* The author hosts the Library and Information Science Laboratory of the Graduate School of Education. He is also affiliated with the Interfaculty Initiative in Information Studies at The University of Tokyo.

1. ISO/TC 37/SC 5 (2015): *ISO 17100:2015. Translation Services - Requirements for translation services*. Geneva: International Organisation for Standardisation. Visited 3 June 2018, <https://www.iso.org/standard/59149.html>.
2. For an overview, see Zakaluk and Samuels (1988) and Feng, Jansche, *et al*. (2010).
3. Khmaladze, Estate V. (1987): *The statistical Analysis of Large Number of Rare Events*. Technical Report MS-R8804. Amsterdam: Centrum Wiskunde & Informatica.
4. R Core Team (2017): *R*. Version 3.4. Visited 15 December 2018, <http://www.r-project.org>.
5. How to calculate the mutual information for a randomised text is explained in Montemurro and Zanette (2010) and Asaishi (2017).
6. The bibliographic information of these texts is as follows:
   L1 and L2: excerpts from Coop, Stephanie (2012): *International Criminal Law from a Gender Perspective*. Doctoral dissertation, unpublished. Tokyo: Aoyama Gakuin University.
   M1: Abstract of McMahon, Daria M., Vdovenko, Vitaliy Y., Karmaus, Wilfried, *et al*. (2014): Effects of long-term low-level radiation exposure after the Chernobyl catastrophe on immunoglobulins in children residing in contaminated areas: prospective and cross-sectional studies. *Environmental Health*. 13(1-36). Visited 17 May 2018, <https://doi.org/10.1186/1476-069X-13-36>.

M2: Abstract of Stepanova, Eugenia, Karmaus, Wilfried, Naboka, Marina, et al. (2008): Exposure from the Chernobyl accident had adverse effects on erythrocytes, leukocytes, and, platelets in children in the Narodichesky region, Ukraine: A 6-year follow-up study. *Environmental Health*. 7(1-21). Visited 10 May 2018, <https://doi.org/10.1186/1476-069X-7-21>.

P1: Podur, Justin (25 Septermber 2015): The most dangerous moment in Colombia's peace talks. *ZNet*. Visited 24 April 2018, <https://zcomm.org/znetarticle/the-most-dangerous-moment-in-colombias-peace-talks/>.

P2: Podur, Justin (11 February 2016): 'Paz Colombia': the latest US attempt to control Colombia? *ZNet*. Visited 25 April 2018, <https://zcomm.org/znetarticle/paz-colombia-the-latest-us-attempt-to-control-colombia/>.

S1: Abstract of Hu, Yan, Bürgmann, Roland, Banerjee, Paramesh, *et al.* (2016): Asthenosphere rheology inferred from observations of the 2012 Indian Ocean earthquake. *Nature*. 538:368-372.

S2: Abstract of Masuti, Sagar, Barbot, Sylvain D., Karato, Shun-ichiro, et al. (2016) Upper-mantle water stratification inferred from observations of the 2012 Indian Ocean earthquake. *Nature*. 538:373-377.

T1: Meadows, Chris (June 10 2016): Could Apple app subscription model portend changes to e-book royalties? *TeleRead*. Visited 30 April 2018, <http://teleread.com/could-apple-app-subscription-model-portend-changes-to-e-book-royalties/>.

T2: Meadows, Chris (June 14 2016): iBooks Editions: Too little, too late? *TeleRead*. Visited 1 May 2018, <http://teleread.com/ibooks-editions-too-little-too-late/>.

W1: Wikipedia contributors (Last update: 15 July 2018): Philosophy of mind. *Wikipedia, The Free Encyclopedia*. Visited 17 July 2018, <https://en.wikipedia.org/w/index.php?title=Philosophy_of_mind&oldid=880290394>.

W2: Wikipedia contributors (Last update: 14 July 2018): Mind-body problem. *Wikipedia, The Free Encyclopedia*. Visited 17 July 2018, <https://en.wikipedia.org/w/index.php?title=Mind%E2%80%93body_problem&oldid=882415118>.

REFERENCES

Ahmad, Khurshid and Rogers, Margaret (2001): Corpus linguistics and terminology extraction. *In*: Sue Ellen Wright and Gerhard Budin, eds. *Handbook of Terminology Management*. Vol. 2. Amsterdam/Philadelphia: John Benjamins, 725-760.

Aizawa, Akiko (2000): An information-theoretic perspective of tf-idf measures. *Information Processing and Management*. 39(1):45-65.

Asaishi, Takuma (2017): *An Informetric Analysis of the Arrangement of Knowledge in High-school Science Textbooks*. Doctoral dissertation, unpublished. Tokyo: The University of Tokyo.

Asaishi, Takuma and Kageura, Kyo (2016): Growth of the terminological networks in junior-high and high school textbooks. *In*: Fahad Khan, Špela Vintar, Pilar León Araúz, *et al*., eds. *LangOnto2 + TermiKS Proceedings*. (LangOnto2 + TermiKS: Joint Second Workshop on Language and Ontology & Terminology and Knowledge Structures, Portorož, 23 May 2016). Paris: European Language Resources Association, 30-37.

Baayen, R. Harald (2001): *Word Frequency Distributions*. Dordrecht: Kluwer.

Bowker, Lynne (2015): Terminology and translation. *In*: Hendrik. J. Kockaert and Frieda. Steurs, eds. *Handbook of Terminology*. Vol. 1. Amsterdam/Philadelphia: John Benjamins, 304-323.

Byrne, Jody (2006): *Technical Translation: Usability Strategies for Translating Technical Documentation*. Dordrecht: Springer.

Cabré, Maria Teresa (2010): Terminology and translation. *In*: Yves Gambier and Luc van Doorslaer, eds. *Handbook of Translation Studies*. Vol. 1. Amsterdam/Philadelphia: John Benjamins, 356-365.

Efron, Bradley and Thisted, Ronald (1976): Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*. 63(3):435-447.

Evert, Stefan (2004): A simple LNRE model for random character sequences. *In*: Gérald Purnelle, Cédrick Fairon, and Anne Dister, eds. *Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*. (JADT2004: 7es Journées

internationales d'Analyse statistique des Données Textuelles, Louvain-la-Neuve, 10-12 March 2004). Vol. I. Louvain-la-Neuve: Presses universitaires de Louvain, 411-422.

Evert, Stefan and Baroni, Marco (2007): zipfR: Word frequency distributions in R. *In*: Annie Zaenen and Antal van den Bosch, eds. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. (ACL'07: 45th Annual Meeting of the Association for Computational Linguistics, Prague, 25-27 June 2007). Stroudsburg: Association for Computational Linguistics, 29-32.

Felber, Helmut (1984): *Terminology Manual*. Paris: Unesco/Inforterm.

Feng, Lijun, Jansche, Martin, Huenerfauth, Matt, *et al*. (2010): Comparison of features for automatic readability assessment. (Coling 2010, Beijing, 23-27 August, 2010) 276-284.

Gotti, Maurizio and Šarčević, Susan (2006): Introduction. *In*: Maurizio Gotti and Susan Šarčević, eds. *Insights into Specialized Translation*. Bern: Peter Lang, 9-24.

Gries, Stefan Th. (2009): *Quantitative Corpus Linguistics with R: A Practical Introduction*. London/New York: Routledge.

Herrera, Juan P. and Purry, Pedro A. (2008): Statistical keyword detection in literary corpora. *European Physical Journal*. B63:135-146.

Heylen, Kris and de Hertog, Dirk (2015): Automatic term extraction. *In*: Hendrik. J. Kockaert and Frieda Steurs, eds. *Handbook of Terminology*. Vol. 1. Amsterdam/Philadelphia: John Benjamins, 203-221.

Jones, Karen (1995): Readability of textbooks for technology education. *Technology Teacher*. 55:28-32.

Kageura, Kyo (2012): *The Quantitative Analysis of the Dynamics and Structure of Terminologies*. Amsterdam/Philadelphia: John Benjamins.

Kageura, Kyo and Kikui, Genichiro (2006): A self-referring quantitative evaluation of the ATR basic travel expression corpus (BTEC). *In*: Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, *et al*., eds. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. (LREC 2006: Fifth International Conference on Language Resources and Evaluation, Genoa, 24-26 May 2006). Genoa: European Language Resources Association, 1945-1950.

Kageura, Kyo and Umino, Bin (1996): Methods of automatic term recognition. *Terminology*. 3(2):259-289.

Mandelbrot, Benoit (1953): An information theory of the statistical structure of language. *In*: Willis E. Jackson, ed. *Communication Theory*. New York: Academic Press, 503-512.

Manning, Christopher, Raghavan, Prabhakar, and Schütze, Hinrich (2009): *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Manning, Christopher and Schütze, Hinrich (2009): *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Mehri, Ali and Darroneh, Amir (2011): The role of entropy in word ranking. *Physica A*. 390:3157-3163.

Miyata, Rei and Kageura, Kyo (2016): Constructing and evaluating controlled bilingual terminologies. *In*: Patrick Drouin, Natalia Grabar, Thierry Hamon, *et al*., eds. *Proceedings of the 5th International Workshop on Computational Terminology*. (CompuTerm 2016: 5th International Workshop on Computational Terminology, Osaka, 12 December 2016). Osaka: The COLING 2016 Organizing Committee, 83-93.

Montemurro, Marcero and Zanette, Damián (2010): Towards the quantification of semantic information encoded in written language. *Advances in Complex Systems*. 13(2):135-153.

Montero Martinez, Silvia and Faber, Pamela (2009): Terminological competence in translation. *Terminology*. 15(1):88-104.

Nord, Christiane (1988/1991): *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-oriented Text Analysis*. (Translated by Christiane Nord and Penelope Sparrow) Amsterdam/Atlanta: Rodopi.

Oakes, Michael and Ji, Meng, eds. (2012): *Quantitative Methods in Corpus-based Translation Studies*. Amsterdam/Philadelphia: John Benjamins.

Ortuño, Miguel, Carpena, Pedro, Bernaola-Galván, Pedro, *et al*. (2002): Keyword detection in natural languages and DNA. *Europhysics Letters*. 57(5):759-764.

Popescu, Ioan-Ioviz (2009): *Word Frequency Studies*. Berlin: Mouton de Gruyter.

Pearson, Jennifer (1998): *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

Rey, Alain (1995): *Essays on Terminology*. (Translated by Juan C. Sager) Amsterdam/Philadelphia: John Benjamins.

Rogers, Margaret (2008): Terminological equivalence: Probability and consistency in technical translation. *In*: Heidrun Gerzymisch-Arbogast, Gerhard Budin, and Gertrud Hofer, eds. *LSP Translation Scenarios: Selected Contributions to the EU Marie Curie Conference Vienna 2007*. (MuTra 2007: LSP translation scenarios, Vienna, 30 April-4 May, 2007). *MuTra*. 02:101-108.

Sager, Juan (1990): *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Sager, Juan, Dungworth, David, and McDonald, Peter (1980): *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Oscar Brandstetter.

Stamatatos, Efstathios (2009): A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*. 60(3):538-556.

Temmerman, Rita (2000): *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins.

Tuldava, Juhan (1995): *Methods in Quantitative Linguistics*. Trier: Wissenchaftlicher Verlag Trier.

Wright, Sue Ellen and Wright, Leland D. Jr. (1997): Terminology management for technical translation. *In*: Sue Ellen Wright and Gerhard Budin, eds. *Handbook of Terminology Management*. Vol. 1. Amsterdam/Philadelphia: John Benjamins, 147-159.

Yang, Zhen, Lei, Jianjun, Fan, Kefeng, and Lai, Yingxu (2011): Keyword extraction by entropy difference between the intrinsic and extrinsic mode. *Physica A*. 392:4523-4531.

Yule, George (1944): *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Zakaluk, Beverley and Samuels, S. Jay, eds. (1988): *Readability: Its Past, Present and Future*. Newark: The International Reading Association.

Zhou, Hongding and Slator, Gary (2003): A metric to search for relevant words. *Physica A*. 329:309-327.

Zipf, George (1935): *The Psycho-biology of Language*. Boston: Houghton Mifflin.