Chapter 4

Applying computer-assisted coreferential analysis to a study of terminological variation in multilingual parallel corpora

Koen Kerremans
Vrije Universiteit Brussel

Coreferential analysis involves identifying linguistic items (usually both lexical and grammatical items) that denote the same referent in a given text. To be able to study such coreferential items, each item first needs to be indexed or annotated according to a referent's corresponding identification code or label. Linguistic items that are identified as 'coreferential' can be represented in a coreferential chain, i.e. a list of coreferential items extracted from the text in which the order of the items in the text is retained. We will discuss some of the benefits of applying coreferential analysis to a study of intra- and interlingual terminological variation in multilingual parallel corpora. Intralingual terminological variation refers to the different ways in which specialised knowledge can be expressed by means of terminological units (both single and multiword units) in a collection of source texts. Interlingual variation pertains to the different ways in which these source language terms are translated into the languages of the target texts. In this contribution, I will focus on how the method of coreferential analysis was used in a comparative study of (intra- and interlingual) terminological variation in original texts (i.e. the source texts) and their translations (i.e. the target texts). I will present a semi-automatic method to support the manual identification of intralingual terminological variants based on coreferential analysis. We will discuss how data resulting from coreferential analysis can be used to quantitatively compare terminological variation in source and target texts. Finally, I will present a new type of translation resource in which terminological variants in the source language are represented as a network of coreferential links.

Koen Kerremans. 2017. Applying computer-assisted coreferential analysis to a study of terminological variation in multilingual parallel corpora. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 45–68. Berlin: Language Science Press. DOI:10.5281/zenodo.814464

1 Introduction

The work presented in this contribution further builds on a research study that focused on how terms and equivalents recorded in multilingual terminological databases can be extended with terminological variants and their translations retrieved from English source texts and their translations into French and Dutch (Kerremans 2012). First, a distinction needs to be made between intralingual (terminological) variation and interlingual variation. The former refers to different ways in which specialised knowledge can be expressed by means of terms in a collection of source texts. Interlingual variation pertains to a study of the different ways in which these source language terms were translated into the languages of the target texts.

In many terminology approaches, terminological variants within and across languages are identified on the basis of semantic and/or linguistic criteria (Carreño Cruz 2008; Fernández Silva 2010). Given the fact that the general aim of the study reported by Kerremans (2012) was to examine how and to what extent patterns of variation in source texts are reflected in the translations, I decided to apply coreferential analysis to the study of (intralingual) terminological variation in the source texts and contrastive analysis to the study of interlingual variation. Our approach based on these perspectives of analysis is motivated by the fact that in order to acquire an understanding about the unit of specialised knowledge or 'unit of understanding' (Temmerman 2000)[1] that needs to be translated, translators first analyse the different ways in which this unit is expressed in the source text, how its meaning is developed in the text (i.e. the textual perspective) and how it can be rendered in the target language (i.e. the contrastive perspective). The combination of coreferential and contrastive methods of analysis allows us to retrieve a list of terminological units for a preselected set of units of understanding in the source texts and to compare this list to the equivalents of each terminological unit in the target texts.

In text-linguistic approaches to the study of terminology (Collet 2004), it has been advocated that terms function as cohesive devices in a text in the sense that they contribute to the reader's general understanding of the text and, in particular, of the units of understanding (Temmerman 2000). As a result of this, the occurrence of terminological variants in a given text is also functional in the sense that these variants allow authors to express their different ways of 'looking' at the same units of understanding (Cabrè 2008; Freixa, Fernández Silva & Cabrè 2008; Fernández Silva 2010).

[1] In (Temmerman 2000), the term 'unit of understanding' is used instead of 'concept' to emphasise the prototypical structure of specialised knowledge.

Within text-linguistic studies, coreferential analysis is a method for linguistic analysis that is used to study patterns of cohesion in a text (Section 2). The purpose of this contribution is to discuss some of the benefits of applying coreferential analysis to a study of intra- and interlingual terminological variation in multilingual parallel corpora (Section 3). My focus will be on three topics in particular:

1. the possibility to support the process of identifying terminological variants as coreferential items by means of a semi-automatic method (see Section 4);
2. the possibility to carry out quantitative comparisons of terminological variants that are identified on the basis of coreferential analysis (see Section 5);
3. the possibility to create a new type of translation resource in which terminological variants in the source language are represented as a network of coreferential links (see Section 6).

By focusing on these three topics in particular, I hope to provide research ideas for future (quantitative and qualitative) studies adopting a textual perspective to terminological variation (see Section 7).

2 Research background

In this section, I want to make clear how terminological variation is defined in the present study (see Section 2.1). Given the fact that I adopt a textual perspective to the study of this phenomenon (see previous section), I want to briefly describe what this perspective involves and how coreferential analysis fits within this perspective (see Section 2.2).

2.1 Terminological variation as the object of study

A study of terminological variation can theoretically pertain to any set of terms in a domain's specialised discourse. In practice, boundaries will need to be drawn in order to limit the scope of the study to a scalable subset of data. According to Daille (2005), these boundaries can be determined by the potential use of the results of the study in various applications (e.g. information retrieval, machine-aided text indexing, scientific and technology watch and controlled terminology for computer-assisted translation systems), the computer techniques

involved in studying the phenomenon and/or the types of language data (mono-/bi-/multilingual data). The application-oriented view explains why a definition of the phenomenon in one study of terminological variation cannot simply be applied to another study.

Based on a review of earlier studies of terminological variation, Cea & Montiel-Ponsoda (2012) present a typology of term variants that is based on a three-fold structure:

1. The first group encompasses a group of synonyms or terminological units that refer to an identical concept. The types of term variants that enter this group are graphical and orthographical variants (e.g. 'Kyoto-protocol' vs. 'Kyoto protocol'), inflectional variants (e.g. 'introduction' and 'introductions') or morpho-syntactic variants ('greenhouse gas emissions' and 'emissions of greenhouse gases').

2. The second group of variants covers partial synonyms or terminological units that highlight different aspects of the same concept. To this group belong stylistic or connotative variants (e.g. 'recession' vs. 'r-word'), diachronic variants (e.g. 'tuberculosis' and 'phthisis'), dialectical variants ('gasoline' vs. 'petrol'), pragmatic or register variants (e.g. 'swine flu' vs. 'pig flu' vs 'Mexican pandemic flu' vs. 'H1N1') and explicative variants ('immigration law' vs. 'law for regulating and controlling immigration'). Examples of these types have been studied in different fields (Temmerman 1997; Resche 2000; Fernández Silva 2010).

3. The third group of variants covers terminological units that show formal similarities but refer to different concepts Daille et al. (1996); Arlin et al. (2006); Bowker & Hawkins (2006); Depierre (2007). Examples are terms showing lexical similarities (e.g. 'Kyoto-protocol' vs. 'Kyoto mechanism') or morphological similarities (e.g. 'biodiversity' vs. 'biosphere' vs. 'biology').

In my study, terminological variation pertains to the first two groups of variants discussed by Cea & Montiel-Ponsoda (2012). It was stated earlier (see Section 1) that as far as intralingual terminological variation is concerned, I applied coreferential analysis to study this phenomenon in a collection of source texts. This implies a textual perspective to the study of terminological variation that I want to briefly discuss in the next section before I explain how the method of coreferential analysis was carried out in my study (see Section 3).

2.2 A textual perspective applied to terminological variation

Within the textual perspective, a distinction needs to be made between text coherence and text cohesion. Based on an extensive review of literature addressing these two topics, Tanskanen (2006: 7) notes that there is a general consensus to define cohesion and coherence as follows:

> "Cohesion refers to the grammatical and lexical elements on the surface of a text which can form connections between parts of the text. Coherence, on the other hand, resides not in the text, but is rather the outcome of a dialogue between the text and its listener or reader. Although cohesion and coherence can thus be kept separate, they are not mutually exclusive, since cohesive elements have a role to play in the dialogue."

Cohesion and coherence contribute to the general texture within a text. In other words, they are a set of characteristics that allows the text to function as a whole. Cohesion is generally regarded as a text internal property, whereas coherence is not. The latter can only be attributed to the text by the reader who is thought to use background knowledge during the interpretation process of the text. This allows the reader to create correlates between the text and the outside world. This knowledge "encompasses beliefs and assumptions about the world as well as language-related knowledge, i.e. knowledge about grammar and about words and their meanings but also knowledge about how texts function" (Collet 2004: 104). Given the fact that the focus of this study is on terminological variation in texts, I will only be concerned with text cohesion.

Cohesion as a text internal property is created on the basis of connected text fragments that allow meaning to pass from one text fragment to another, thus establishing cohesive chains within the text. Collet (2004) describes these as "chains of text fragments that refer to the same concrete or abstract reality" and "which can be obtained with grammatical and lexical means" (ibid.). Halliday & Hasan (1976) propose five types of cohesion: reference, substitution, ellipsis, conjunction and lexical cohesion. Since my study focuses on terms as cohesive devices in texts (see Section 1), I shall only focus on lexical cohesion.

Applied to studies of terminology, lexical cohesion analysis is achieved by means of a selection of a domain's terminology appearing in a text. Halliday & Hasan (1976) distinguish between two types of lexical cohesion: reiteration and collocation. They define the former as a form of lexical cohesion "which involves the repetition of a lexical item, at one end of the scale; the use of a general word to refer back to a lexical item, at the other end of the scale; and a number of things

in between - the use of a synonym, near-synonym, or superordinate" (ibid.: 278). Collocation occurs between any pair of lexical items "that stand to each other in some recognizable lexico-semantic (word meaning) relation" (ibid: 285). In other words, the 'collocation' refers to "an associative meaning relationship between regularly co-occurring lexical items" (Tanskanen 2006: 12).

In the present study, terminological variation is clearly seen as the result of a process of reiteration whereby the author of a text uses the same or different terminological units to express the same unit of understanding. In this perspective, coreferential analysis is a technique that is suitable for identifying those linguistic items that refer to the same unit of understanding in a text. To be able to study such coreferential items, each item first needs to be indexed or annotated according to a referent's corresponding identification code or label. Linguistic items that are identified as 'coreferential' can be represented in a coreferential chain, i.e. a list of coreferential items extracted from the text in which the order of the items in the text is retained.

Rogers (2007) shows how the technique of coreferential analysis can be used to study patterns of terminological equivalence between source and target texts. By presenting terminological variants as coreferents in lexical chains she is able to compare the use of terms in establishing cohesive ties in a German technical text and its translations into English and French. Before I illustrate on the basis of examples from my own study how this method is carried out, I will first briefly present in the next section the research design of the case study presented by Kerremans (2012), which forms the basis for the present study. This will allow us to motivate the particular choices that were made with respect to the method of analysis.

3 Intra- and interlingual variation in parallel texts

The general aim of the study described in Kerremans (2012) was to try to understand how translators of specialised texts tend to deal with terminological variation in texts that need to be translated (i.e. source texts). For instance, a topic such as the rise in the average temperature of the earth's surface can be referred to in English as 'global warming', 'greenhouse effect' or 'hothouse effect'. By comparing such terms in English source texts with their translations in Dutch and French versions of these texts (i.e. target texts), the overall aim of this study was to acquire a better insight into various ways of translating English environmental terminology into Dutch and French.

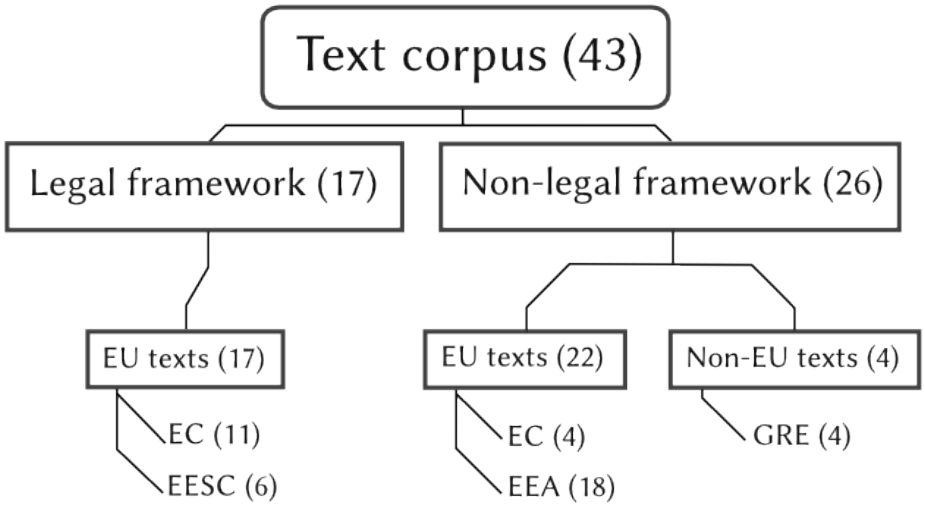

Figure 1: Classification of texts

The corpus created for this study is comprised of 43 texts. Each text is available in three language versions - English, French and Dutch - which means that in total 129 texts were used to study patterns of intra- and interlingual variation. All the texts in the corpus were originally written in English and translated into French and Dutch. The texts dealt with environmental topics, such as biodiversity loss, climate change, invasive species and environmental pollution. Texts were collected from different organisations (mainly EU institutions) and written registers (e.g. EU directives, information brochures, etc.) in order to study variation in relation to different situational parameters, such as text source, text framework (see Section 6). Figure 1 shows how the texts in the corpus were classified according to different text perspectives.

First of all, a distinction is made between 17 texts (69,647 words in the English versions) belonging to the legal framework (e.g. EC communications, green papers and staff working documents, EESC opinions) and 26 texts (39,183 words in the English versions) that do not belong to this framework (e.g. fact sheets and booklets). Within the first category, only EU texts were added to the corpus. Within the second category, a further distinction was made between 22 EU texts and 4 non-EU texts. Apart from these two text dimensions, texts were also classified according to the institution responsible for the trans-lation and publication of the texts: the European Economic and Social Committee (EESC), the European Commission (EC), the European Environment Agency (EEA) and, finally, Green-

facts (GRE), a non-profit organisation that summarises and translates scientific publications on health and environmental issues for the general public. [2]

As was mentioned in the beginning (see Section 1), the research data (i.e. both intra- and interlingual variation) were collected from this corpus by applying both coreferential and contrastive analyses. In total, approximately 9,100 terminological variants were extracted from the English source texts on the basis of coreferential analysis. By applying a contrastive perspective, the translation equivalents of these English variants were retrieved from the French and Dutch target texts. The combination of an English term and its translation in either French or Dutch (i.e. a Translation Unit or TU), is stored in a separate database. The result was a database of approximately 18,200 TUs (English-French; English-Dutch).

Quantitative comparisons of these translation units were carried out in subsequent phases of the project. Each TU is comprised of a term in the source language (i.e. English), its corresponding equivalent that was retrieved from the target text in combination with additional contextual information: i.e. a specification of the unit of understanding to which the source language term refers as well as information about specific properties of the text from which the TU was retrieved.

Given the fact that the focus of this contribution is on coreferential analysis, it will be briefly illustrated by means of the example in Figure 2 how this particular analysis was carried out.

The figure contains an annotation scheme featuring 10 cluster labels and a text sample taken from a European Commission Staff Working document (European Communities 2008: 2). Cluster labels are ad-hoc labels that were created to facilitate the annotation of English terminological variants as coreferential items. Each cluster label represents a particular unit of understanding (see Section 1). For instance, the cluster label INVASIVE_ALIEN_SPECIES represents the unit of understanding (or conceptual category) that can be described as 'species that enter a new habitat and threaten the endemic fauna and/or flora'. Terminological variants that are annotated according to this label will appear in the lexical chain or 'cluster' of terms denoting the same unit of understanding in the text (see Table 1). For instance, the lexical chain drawn from the text sample in Figure 2 for the unit of understanding INVASIVE_ALIEN_SPECIES is: invasive alien species – IAS – invasive species – IS – IS – IS – invader.

[2] Texts from the European Economic and Social Committee (EESC) and the Committee of the Regions (COR) were classified according to one category EESC because texts from both institutions are translated by the same translation department.

Annotation scheme	Text sample
ALIEN_SPECIES BIOCONTROL BIODIVERSITY BIO-INVASION ECOSYSTEM FORESTRY INTRODUCTION INVASIVE_ALIEN_SPECIES MEA SPREAD	[Invasive Alien Species]$_{\text{INVASIVE\_ALIEN\_SPECIES}}$" are [alien species]$_{\text{ALIEN\_SPECIES}}$ whose [introduction]$_{\text{INTRODUCTION}}$ and/or [spread]$_{\text{SPREAD}}$ threaten [biological diversity]$_{\text{BIODIVERSITY}}$ [...]. The [Millennium Ecosystem Assessment]$_{\text{MEA}}$ revealed that [IAS] $_{\text{INVASIVE\_ALIEN\_SPECIES}}$ impact on all [ecosystems]$_{\text{ECOSYSTEM}}$ [...]. The problem of [biological invasions]$_{\text{BIO-INVASION}}$ is growing rapidly as a result of increased trade activities. [Invasive species] $_{\text{INVASIVE\_ALIEN\_SPECIES}}$ ([IS] $_{\text{INVASIVE\_ALIEN\_SPECIES}}$) negatively affect [biodiversity]$_{\text{BIODIVERSITY}}$ [...]. [IS] $_{\text{INVASIVE\_ALIEN\_SPECIES}}$ can cause congestion in waterways, damage to [forestry]$_{\text{FORESTRY}}$, crops and buildings and damage in urban areas. The costs of preventing, controlling and/or eradicating [IS]$_{\text{INVASIVE\_ALIEN\_SPECIES}}$ and the environmental and economic damage are significant. The costs of [control]$_{\text{BIOCONTROL}}$, although lower than the costs of continued damage by the [invader]$_{\text{INVASIVE\_ALIEN\_SPECIES}}$, are often high.

Figure 2: Example illustrating coreferential analysis

Table 1: Results of the coreferential analysis

Cluster label	Lexical chain
INVASIVE_ALIEN_SPECIES	invasive alien species - IAS - Invasive species - IS - IS - IS - invader
ALIEN_SPECIES	alien species
INTRODUCTION	introduction
SPREAD	spread
BIODIVERSITY	biological diversity - biodiversity
MEA	Millennium Ecosystem Assessment
ECOSYSTEM	ecosystems
BIO-INVASION	biological invasions
FORESTRY	forestry
BIOCONTROL	control

Co-referential analysis focuses on reformulation procedures, which according to Ciapuscio (2003: 212), are procedures defined mainly on the basis of structural criteria, such as "the rewinding loop in speech, the resumption of an idea that has already been verbalized, which is linguistically realised in the two-part structure "referential expression" + "treatment expression", both expressions usually being linked with markers." The first term ('Invasive Alien Species') which introduces the unit of understanding INVASIVE_ALIEN_SPECIES in the text sample (see 2) is called the 'referential expression'. It represents the perspective from which the referent should be perceived. This is the reason why all coreferential expressions in Figure 2 are annotated according to the cluster label INVASIVE_ALIEN_SPECIES. The expressions that follow the referential expression are called treatment expressions because they reveal a new aspect of the referent. The choice for a particular cluster label is determined by the referential expression, not by the treatment expression. For instance, the term 'alien species' may be annotated as ALIEN_SPECIES or as INVASIVE_ALIEN_SPECIES, depending on whether the term occurs as referential expression or treatment expression (i.e. shortened form of the term invasive alien species).

Coreferential analysis in my study was guided by the following rules:

- Every term candidate had to be a nominal pattern in order to have a common basis for comparing intralingual variants. The focus on nominal patterns makes sense in the context of terminology work, in which "the predominance of nouns is an incontestable phenomenon" (Bae 2006: 19). According to L'Homme (2003: 404) this focus on nominal patterns can be justified by the fact that specialised knowledge is usually "represented by terms that refer to entities (concrete objects, substances, artifacts, animates, etc.), and that entities are linguistically expressed by nouns."

- Every term candidate that is part of a linguistic construction that refers to a different unit of understanding should not be annotated. For instance, even though the pattern 'alien species' occurs two times in the text sample in Figure 2, only the second occurrence is marked with the corresponding ALIEN_SPECIES. This is because in the first occurrence, the term is part of the linguistic pattern 'invasive alien species' which refers to the unit of understanding INVASIVE_ALIEN_SPECIES.

- Every term candidate that is not part of a linguistic construction that refers to a different unit of understanding should be annotated. This rule applied to term candidates that are not part of a nominal construction - such as

'invasive alien species', 'invasive species' or 'biological diversity' (see 2) - or term candidates that are part of a nominal construction that did not refer to a different unit of understanding in my dataset. The term candidate 'control', for instance, was annotated as 'BIOCONTROL'. The term candidate appears in the nominal construction 'the costs of control', which did not refer to a different unit of understanding in my study.

- Every article or pronoun preceding a term candidate should be left out. For instance, in the nominal constituent 'The Millennium Ecosystem Assessment' (see 2), the article preceding the term candidate was not taken up in the analysis.
- All term candidates that are linked to one another in the same nominal pattern by means of coordinating conjunctions should be annotated separately. For instance, the pattern 'introduction and/or spread' features two different units of understanding in my dataset: resp. INTRODUCTION and SPREAD. More complex patterns to annotate were conjunctive patterns featuring different modifiers linked to one head. Consider for instance the text string 'invasive and alien species' which comprises two term variants ('invasive species' and 'alien species') that should be classified according to two different clusters: INVASIVE_ALIEN_SPECIES and ALIEN_SPECIES. The second term candidate in this pattern (i.e. 'alien species') does not pose any problem. The occurrence can be immediately extracted from the text without any modifications required. The first term candidate (i.e. 'invasive species'), however, could not be directly extracted as it is interrupted by the conjunction word 'and' and the adjective 'alien'. To be able to annotate this term candidate as occurrence of the unit of understanding INVASIVE_ALIEN_SPECIES and to add the correct form 'invasive species' to a separate database containing the research data, a distinction had to be made between occurrences and base forms. The occurrence refers to the English term variant as it appeared in the corpus. The base form is a 'cleaned' version of the occurrence in which possible irrelevant words in multiword terms are deleted. In the example of 'invasive and alien species', for instance, the base form of this pattern referring to the cluster INVASIVE_ALIEN_SPECIES is 'invasive species'. It should be noted that results derived from the quantitative analyses of intra- and interlingual variants in the corpus are based on the comparisons of base forms only (Section 5).

In Section 1, I mentioned that the purpose of this article is to discuss some of the benefits of applying the aforementioned method to a study of terminological

variation in multilingual parallel corpora. In the remainder of this contribution, I will focus on the possibility to support the manual effort by means of automated procedures (Section 4), the possibility to carry out quantitative comparison of terminological variants in lexical chains (see Section 5) and, finally, the possibility to create a new type of translation resource in which terminological variants in the source language are represented as a network of coreferential links (see Section 6).

4 Computer-assisted coreferential analysis

A major drawback of the method outlined above is the fact that it is very difficult to apply if the work is only carried out manually. During the coreferential analysis of the source texts, 241 cluster labels needed to be taken into consideration in our study. Given the fact that the process of annotating or 'labeling' terminological variants as 'coreferential' involves performing manual actions which are to a certain degree repetitive and predictable, I developed a semi-automatic method to support this labour-intensive process.

Before I outline this method, it should be noted that different approaches have been proposed for automatically extracting intralingual terminological variation from texts. Some approaches are based on the search for contexts that contain predefined sets of text-internal markers, called Knowledge Patterns or KPs. In literature, such patterns are often used to extract two term candidates linked by a specific semantic relation. For a survey of such approaches, see Auger & Barrière (2008). In other approaches, terminological variants are identified on the basis of distributional measures. The basic idea in these approaches is that the more distributionally similar two term candidates are, the more likely that they can be used interchangeably in linguistic contexts (Weeds & Marcu 2005; Rychlý & Kilgarriff 2007; Shimizu et al. 2008; Kazama et al. 2010). A major disadvantage of approaches based on distributional measures is the difficulty to understand the types of semantic relations (e.g. synonymy, hyperonymy, antonymy, etc.) that can be inferred from the resulting clusters of words or terms (Budanitsky & Hirst 2006; Heylen, Peirsman & Speelman 2008; Peirsman, Heylen & Speelman 2008).

In order to make sure that, for the preselected set of units of understanding, all English terminological variants and their translations into French and Dutch would be retrieved from the trilingual corpus (see Section 3), while retaining the order of appearance of each variant in the texts, I decided to support my manual coreferential and contrastive methods of analysis by means of automated procedures. This semi-automatic approach allows us to ensure completeness,

accuracy and consistency in the data obtained. The automated procedures are implemented in a script that was written in the Perl programming language[3].

Given the scope of this study, I shall only focus on the computer-assisted method supporting the manual identification of coreferential terminological variants in the English source texts. The purpose of this method is threefold: (1) to support the identification of terminological variants that are coreferentially linked to a common unit of understanding, (2) to annotate these variants according to a common cluster label (see Section 3) and (3) to extract these variants from the text and store them as lexical chains in a separate database.

It should be noted that prior to this method, each source text in the corpus needs to be aligned with its corresponding text(s) in the target language(s). After that, the script developed to support coreferential analysis reads every text segment (usually corresponding with a sentence in the text) one after the other and carries out a number of tasks. For each term variant that is manually selected in a text segment, the script will first suggest possible matching cluster labels, based on term variants that were manually entered in a previous stage. If no matching clusters were found, the proper cluster label needs to be specified by the user.

After that, the new term variant and its corresponding cluster label are stored in a dataset of 'Clusters'. Whenever the term variant is found in the subsequent text segments, it is automatically identified as a term candidate and its corresponding cluster label is presented to the user. In case of term variants that are already 'known' to the system, the user simply needs to confirm or reject the suggestions made by the system.

The computer-assisted method relies on three resources during the analysis of coreferential terminological variants in the source texts: i.e. 'Clusters', 'Filtering rules' and a 'Dictionary' (see Figure 3).

The function of each resource is explained as follows:

- 'Clusters': a dataset of all the cluster labels (see above) and the term variants already encountered in previous texts. The dataset is used to automatically identify and cluster term variants that were previously encountered during coreferential analysis. This dataset continuously grows as more variants are retrieved from texts.
- 'Filtering rules': a list of rules comparable to a stoplist. It contains patterns that should be ignored during the search for term candidates. As the search for term candidates was case-insensitive, for instance, the term candidate 'IS' pointing to the unit of understanding INVASIVE_ALIEN_SPECIES,

[3] https://www.perl.org/

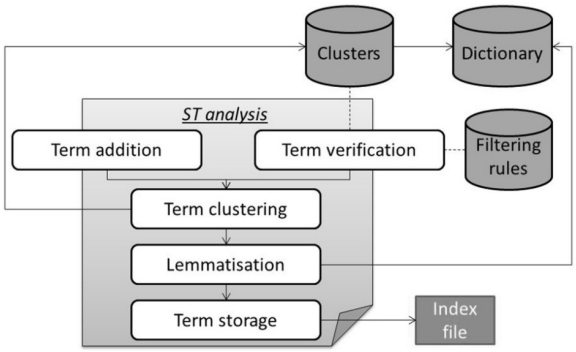

Figure 3: Computer-assisted coreferential analysis

more frequently occurred in the corpus as the third person singular of the verb 'to be'. Filtering rules specifying common patterns in which this form appears as a verb were necessary to exclude the irrelevant occurrences during the analysis of the source texts. Another example is for instance the term candidate 'community' referring to the unit of understanding BIOLOGICAL_COMMUNITY. Filtering rules were created to disregard occurrences of this string in patterns like 'scientific community' or 'economic community' which also frequently occurred in my corpus.

- 'Dictionary': a resource comprised of all occurrences retrieved from the source texts, together with their lemmatised forms. The distinction between lemmatised forms and actual occurrences was necessary to be able to deal with frequently encountered discontinuous multiword expressions such as the term 'control of invasive species' in the string 'control and prevention of invasive species'. Term occurrences were stored in the 'Clusters' dataset (see Figure 5), whereas lemmatised forms were stored in the dictionary.

The semi-automatic method is implemented in such a way that the three aforementioned resources are updated and expanded with new data, any time during the analysis. As a result, the time spent on manually extracting the lexical chains from the source texts is considerably reduced as the analysis proceeds.

Figure 3 also visualises the different semi-automated steps to add term variants to an index file, together with information about their position in the source text and their corresponding cluster labels. This index file is used in a later phase of the project to semi-automatically retrieve the translation equivalents from the

aligned target texts. The semi-automated steps supporting coreferential analysis are:

- 'Term addition': a semi-automated process that can be broken down into the following steps: a) in every text segment, a new term variant is manually highlighted, b) candidates of cluster labels are automatically proposed in the 'Term clustering' procedure and c) the new term variant is automatically added to the 'Clusters' dataset.
- 'Term verification': a semi-automated process whereby text strings corresponding to term candidates in the dataset of 'Clusters' are automatically selected as term variants. After manual validation, potentially relevant cluster labels are looked up in the dataset of clusters on the basis of the 'Term clustering' procedure (see the next step).
- 'Term clustering': a semi-automated process for assigning a proper cluster label to an already familiar term variant. Candidates of cluster labels are automatically proposed based on fuzzy matching between the new term variant and the variants that are already present in the dataset of 'Clusters'. The proper cluster label is manually selected in case more than one cluster candidate was found. In case only one candidate is found, the automatically proposed cluster can either be manually approved or rejected. In case the term variant should be classified according to a cluster that was not proposed as candidate, this cluster is manually selected from the entire dataset of clusters, after which the 'Clusters' dataset and the 'Dictionary' are updated. Finally, candidates of cluster labels are automatically proposed based on fuzzy matching between the new term variant and the variant clusters (see the 'Lemmatisation' process).
- 'Lemmatisation': a semi-automated process for assigning the correct lemmatised form to a term candidate. Candidates of lemmatised forms are automatically proposed based on fuzzy matching between the new term and the existing lemmatised forms. Next, the proper lemmatised form is manually selected in case more than one candidate was found. In case only one candidate is found, the automatically proposed lemma can either be manually approved or rejected. A lemmatised form has to be manually created in case it does not appear in the dictionary. After this, the dataset of clusters and the dictionary are updated and the validated term is stored in the resulting research data file (see the 'Term storage' procedure).

- 'Term storage': i.e. a semi-automated process for storing the validated occurrences of semantically-structured SL term variants in the aforementioned index file (see above).

The computer-assisted approach proved to be an efficient working method for annotating variants in coreferential chains, especially given the high repetition of frequently occurring patterns in the corpus that needed to be marked with the same cluster labels. Based on this method, it was possible to compile a dataset of approximately 9,100 English term variants retrieved from the corpus of source texts and classified according to a predefined set of 241 cluster labels.

5 Quantitative comparisons

By comparing the lexical chains in the source language with the translations of these chains in French and Dutch that were retrieved from the target texts, it was possible to draw conclusions on the occurrence of intra- and interlingual variation in the corpus.

When studied at the level of the text, interlingual variation occurs when terms appearing in the lexical chains in the source text were not consistently translated into the target texts, such as is the case in the example in Table 2. It can be observed from this table that in the French chain, the terminological choices that were made in the English text are reflected. An exception, for instance, is the translation of the English term 'IAS', which appears in the French translation as the full form 'espèces exotiques envahissantes'.

Table 2: English lexical chain and its translation into French and Dutch

English chain	French translation	Dutch translation
invasive alien species	▶ espèce exotique	~ invasieve uitheemse soort (IUS)
IAS	▶ espèces exotiques envahissantes	~ IUS
invasive species	▶ espèce envahissante	~ invasieve soort
IS	▶ EE	~ IS
IS	▶ EE	~ IS
Invader	▶ Envahisseur	~ IS
invasive species	▶ espèces envahissantes	~ invasieve soort

Quantitative analyses were carried out on the basis of comparisons between the English lexical chains and their translations into French and Dutch. The aim of the quantitative comparisons was to examine to what extent the English lexical chains had an impact on the choices made in the target languages. In order to examine this, I compared the transitions between consecutive lemmatised forms in the different chains. The transition from one form to the other is marked as '0' to indicate that no change occurred (e.g. from 'IS' to 'IS'). Changes in transitions (such as from 'invasive alien species' to 'IAS') are marked by '1'. The result of this analysis is a sequence of the values '1' and '0', which allowed us to create a transition profile for each English lexical chain and its corresponding chain in French and Dutch.

The example in Table 3 shows part of the transition profile for the coreferential chain of INVASIVE_ALIEN_SPECIES in TextID 1 (see Section 3). The transition profile for the coreferential chain is: 1 1 1 0 1 1.

Table 3: Example of a transition profile

Order in the text	**English base forms for** INVASIVE_ALIEN_SPECIES	**Transition**
1	invasive alien species	New
2	IAS	1
3	invasive species	1
4	IS	1
5	IS	0
6	invader	1
7	invasive species	1
	Degree of change:	0,83

The first occurrence 'invasive alien species' is marked as the beginning of a new lexical chain ('New'). The second occurrence 'IAS' differs from the first. The first transition is therefore marked as '1'. The fourth transition is marked as '0' because no change occurred in the transition from occurrence 4 ('IS') to 5 ('IS').

The lexical chain features five changes in the transitions between consecutive lemmatised forms on a total of six transitions. By dividing the first number by the second, a degree of change can be created for each coreferential chain separately. This measure allows for a quantitative comparison of the coreferential patterns in the three languages.

In the example in Table 4, the degrees of change for both English and French are 0,83, whereas for Dutch the value is 0,67. A value close to 1 indicates a high degree of change in the chain, whereas a degree of '0' indicates consistency in the lemmatised forms[4] in the pattern.

Table 4: Quantitative comparison between chains

English lemmatised forms		**French lemmatised forms**		**Dutch lemmatised forms**	
invasive alien species	New	espèce exotique	New	invasief uitheems soort (IUS)	New
IAS	1	espèce exotique envahissant	1	IUS	1
invasive species	1	espèce envahissant	1	invasief soort	1
IS	1	EE	1	IS	1
IS	0	EE	0	IS	0
Invader	1	Envahisseur	1	IS	0
invasive species	1	espèce envahissant	1	invasief soort	1
	0,83		0,83		0,67

Once results of the coreferential profiles and the degrees of change were obtained, two methods were applied for comparing variation in the different languages: one method was based on comparisons of the transition patterns in the three languages, the other on examining possible correlations between the degrees of change (see further).

The results in the first method of comparison were classified according to two possible 'scenarios': either the value was '0' (indicating no change in the transition) or '1' (indicating a change). General results are shown in Figure 4.

In 5,359 of the English cases, no variation was encountered in the transition between lemmatised forms in a chain. This corresponds to 72% of the total cases (n=7,446). A closer examination of this category shows that this pattern of consistency is also reflected in the translations. For instance, for the total set of chains,

[4] Note that each word in a term was lemmatised. In some cases, the lemmatisation of words resulted in multiword terms which were ungrammatical (e.g. *'espèce exotique envahissant' in French or *'invasief uitheems soort' in Dutch). This was necessary to make sure that variation resulting from morphological differences could be excluded from my analysis.

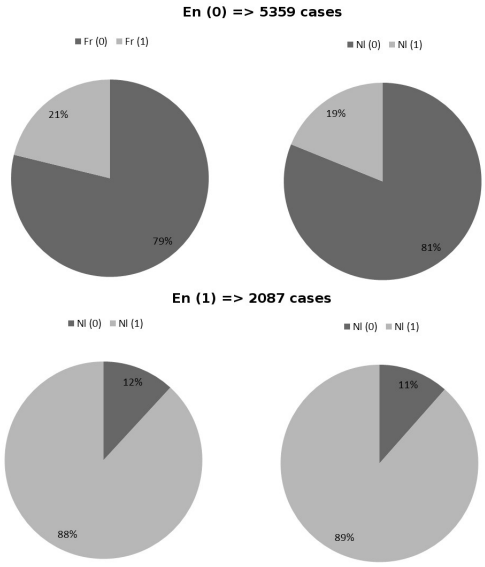

Figure 4: Comparisons of transition patterns

78% of the French cases and 81% of the Dutch cases follow the same pattern as English.

A closer look at the cases that were marked in English as '1' (2,087 cases or 28% of the total cases) shows that the transformations between lemmatised forms in Dutch and French also tend to be marked by this value: 88% of the French cases and 89% of the Dutch cases correspond to the English pattern.

Although these results already give an indication that variation in English coreferential chains is also reflected in the target languages, these results do not show to what extent the degree of variation within a coreferential chain is also reflected in the translations. In the first method, patterns of transition in the three languages are compared on a case by case basis, without taking into consideration the coreferential chain in which the transition takes place.

For this reason, a second type of quantitative comparison was worked out in which the aforementioned degree of variation within each chain was used as a basis for comparison. Given the general hypothesis that the source language has an impact on the choices made in the target language(s), it was hypothesised that the degree of changes in the English coreferential chains would also have a direct impact on the degree of changes in the French and Dutch chains. A

bivariate analysis was conducted in PSPP[5], a free statistical software package, for all subsets in the corpus to determine possible correlations between the degrees of change in the three languages. The results of this analysis are shown in Table 5.

Table 5: Correlations between degrees of variation in coreferential chains

	N	Sig. (1- tailed)	En-Fr	En-Nl
EC (Leg)	456	0.00	0.66	0.61
EC (NLeg)	110	0.00	0.69	0.69
EEA	256	0.00	0.80	0.69
EESC	106	0.00	0.80	0.56
GRE	106	0.00	0.63	0.58
EU	366	0.00	0.76	0.69
Leg	562	0.00	0.69	0.60
Nleg	472	0.00	0.73	0.67
NLeg (EU)	110	0.00	0.69	0.69
Total	1034		0.71	0.63

Positive correlations can be observed in all datasets. The correlations between English and French tend to be stronger than those between English and Dutch. This is particularly the case in the EESC subset which shows a strong correlation between English and French (0,80) and a moderate correlation between English and Dutch (0,56).

6 Coreferential links in a dictionary application

In the previous section, I have shown how results that were partly derived from coreferential analysis can be used for research purposes only, i.e. to compare patterns of variation between source and target texts. In this section, I briefly show how coreferential links can also be used for visualising the relations between intralingual variants in a dictionary application. An example of a prototype visualisation is shown in Figure 5.

The model underlying the representation of variation in Figure 3 is based on the Hallidayan premise that each choice (variant) in a language system acquires its meaning against the background of other choices which could have been made.

[5] http://www.gnu.org/software/pspp/

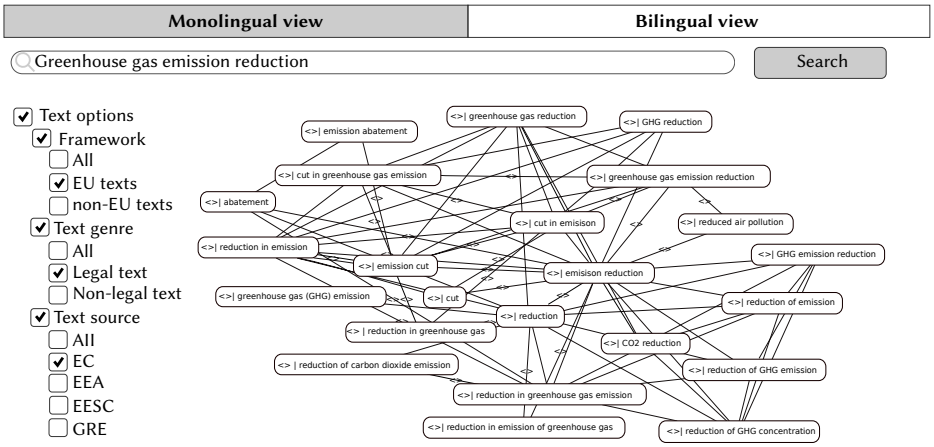

Figure 5: Coreferential links between terminological variants denoting GREENHOUSE_GAS_EMISSION_REDUCTION

These choices are motivated by a complex set of contextual factors which, in Systemic Functional Linguistics, are classified according to the dimensions of Domain, Tenor and Mode (Eggins 2004).

Changing the contextual conditions or options in the model leads to direct changes in the network of terminological options that are shown to the user. Figure 5 shows a network of linguistic (terminological) options for the unit of understanding GREENHOUSE_GAS_EMISSION_REDUCTION in the source language. This network is activated by entering either a SL term appearing in the cluster or the specific cluster label in the search box (at the top). The search query will activate a number of contextual options that are associated with the search. It will also show the results of the search in a graph representation.

Connections in this graph represent the coreferential links between terms appearing in the same texts in the corpus. Selecting or deselecting one or several contextual or linguistic options in the filtering options, will immediately be reflected at the visualisation level. Examples of contextual options in Figure 5 are text options such as those that were mentioned in Section 3 (e.g. text source).

An additional interesting aspect of the graph representation is that it allows for a prototypically-structured visualisation of term variants referring to the same unit of understanding. This means that terms that frequently occurred in all texts have a lot of coreferential connections to other terms in the network. Consequently, these terms will take up a more central position in the network whereas

infrequent patterns will appear more in the periphery. In this way, dictionary users will immediately be able to distinguish between 'core variants' (i.e. variants that are frequently encountered within the selected collection of texts) and 'peripheral variants' (i.e. variants that were only sporadically encountered). Selecting or deselecting certain contextual options can potentially cause term variants to move from the centre to the periphery or vice versa, allowing for dynamic, customised visualisations of semantically-structured term variants.

7 Conclusion

In this contribution, I have discussed how coreferential analysis can be used to identify term variants in a corpus of source texts, how the method can be supported by implementing semi-automatic procedures, how lexical chains in the source language and their translations can form the basis for quantitative comparisons between source and target texts and, finally, how coreferential links between intralingual variants can be represented in a dictionary application.

Coreferential chains can give us more insight in possible patterns of convergence or divergence among language versions of the same document. I therefore intend to further examine cohesive patterns in source and target texts for different reasons. For instance, it can be expected that differences in cohesive patterns will emerge if coreferential analysis is applied to all language versions (instead of the drafted versions only). By comparing the resulting coreferential chains in the different languages, it should be possible to calculate to what extent the target language versions deviate from the source texts in terms of terminological consistency and coreferential patterning. This is for instance valuable information for translators of EU legislation who have to see to it that no deviations occur in language versions of legally-binding texts. Differences in cohesive patterns are thus a possible method for further exploring the notion of anisomorphism in the context of EU translation. Anisomorphism refers to asymmetry in the interlinguistic transfer process, what González-Jover & Gómez (2006: 215)refers to as "the losses and gains that always occur in interlinguistic transfer processes, and which may be taken into account when comparing two different language systems."

Focusing on the coreferential chains in the target language versions will enable us to establish a new type of connection between 'linguistic options' in the source text (not based on the coreferential status of terms in a text but derived from translation similarities). For instance, the English expression 'climate risk' in TextID 6 (see Section 3) was not marked as part of the cluster CLIMATE_IM-

PACT. However, given the fact that in the French text this term was translated as 'conséquences du changement climatique' ('consequences of climate change'), which also appeared in the corpus as the translation of 'climate change impact', a link may be established between the term 'climate risk' and the English cluster of terms denoting CLIMATE_IMPACT:

> *English co-text: "[…] ensuring that long-term infrastructure will be proof to future >>climate risks<< […]"*
>
> French co-text: "[…] soient capables de résister aux >>conséquences du changement climatique<< […]"

Another example in the same text is the English term 'climate-resilient'. This term was not taken up in the cluster CLIMATE_ADAPTATION during the source text analysis but may be linked to this cluster on the basis of its French translation 's'adapter au changement climatique' ('to adapt to climate change')

> *English co-text: "[…] targeted action is needed on building codes and methods, and >>climate-resilient<< crops […]"*
>
> French co-text: "[…] l'élaboration de codes et de méthodes ainsi que la mise en place de cultures pouvant >>s'adapter au changement climatique<< […]"

Although my method proved to be valid for comparing patterns of variation, the time spent in this project on the method of analysis remains a major drawback. Fully automated extraction methods were not used, given the specific research requirements of data accuracy and completeness to be able to compare patterns of variation between the source and target texts. But since the work was characterised by a lot of repetitive tasks (such as selecting and annotating term variants that were previously encountered and thus already known) a combination of automatic procedures and manual verification proved to be efficient. Still, further reflections are necessary to conduct coreferential analyses in a way which seem more efficient and practical from a user perspective. For instance, it will need to be examined how the manual analysis can benefit from an automated co-referential resolution module.

Acknowledgements

The author wishes to thank the reviewers for their valuable comments on an earlier draft.

References

Arlin, Nathalie, Amélie Depierre, Susanne Lervad & Claire Rougemont. 2006. Réflexions sur la variation: étude de cas dans le domaine médical. *LSP and Professional Communication* 6(2). 75–88.

Auger, Alain & Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology* 14(1). 1–19.

Bae, Hee Sook. 2006. Termes adjectivaux en corpus médical coréen: Repérage et désambiguïsation. *Terminology* 12(1). 19–50.

Bowker, Lynne & Shane Hawkins. 2006. Variation in the organization of medical terms: Exploring some motivations for term choice. *Terminology* 12(1). 79–110. http://www.benjamins.com/cgi-bin/t_articles.cgi?bookid=TERM%2012%3A1&artid=7060931, accessed 2008-03-10.

Budanitsky, Alexander & Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1). Cited by 0869, 13–47.

Cabrè, Maria Teresa. 2008. El principio de poliedricidad: La articulación de lo discursivo, lo cognitivo y lo lingüístico en terminología (i). *IBÉRICA* 16. 9–36.

Carreño Cruz, Sahara Iveth. 2008. Characterizing term variation on an English-Spanish parallel corpus. In *Proceedings of Multilingual and Comparative Perspectives in Specialised Language Resources*. Marrakech.

Cea, Guadalupe Aguado-de & Elena Montiel-Ponsoda. 2012. Term variants in ontologies. In *Proceedings of the 30th International Conference of AESLA*, 19–21. Lleida.

Ciapuscio, Guiomar E. 2003. Formulation and Reformulation Procedures in Verbal Interactions between Experts and (Semi-)laypersons. *Discourse Studies* 5(2). 207–233. DOI:10.1177/1461445603005002004

Collet, Tanja. 2004. Esquisse d'une nouvelle microstructure de dictionnaire spécialisé reflétant la variation en discours du terme syntagmatique. *Méta* 49(2). 247–263.

Daille, Béatrice. 2005. Variations and application-oriented terminology engineering. *Terminology* 11(1). 181–197.

Daille, Béatrice, Benoît Habert, Christian Jacquemin & Jean Royauté. 1996. Empirical observation of term variations and principles for their description. *Terminology* 3(2). 197–258.

Depierre, Amélie. 2007. Souvent HAEMA varie …: Les dérivés du grec HAEMA en anglais : étude de cas de variation. *Terminology* 13(2). 155–176. DOI:10.1075/term.13.2.03dep

Eggins, Suzanne. 2004. *Introduction to Systemic Functional Linguistics: 2nd Edition*. London: Continuum International Publishing Group.

European Communities, Commission of the. 2008. *Commission Staff Working Document - Annex to the Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions - Towards an EU strategy on invasive species*.

Fernández Silva, Sabela. 2010. *Variación terminológica y cognición. Factores cognitivos en la denominación del concepto especializado*. Barcelona: Universitat Pompeu Fabra PhD thesis.

Freixa, Judit, Sabela Fernández Silva & Maria Teresa Cabrè. 2008. La multiplicité des chemins dénominatifs. *Meta* 53 (4). 731–747.

González-Jover & Adelina Gómez. 2006. Meaning and anisomorphism in modern lexicography. *Terminology* 12(2). 215–234.

Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman Publishing.

Heylen, Kris, Yves Peirsman & Dirk Speelman. 2008. Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 3243–3249. Marrakech.

Kazama, Jun'ichi, Stijn De Saeger, Kow Kuroda, Masaki Murata & Kentaro Torisawa. 2010. A Bayesian Method for Robust Estimation of Distributional Similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 247–256. Uppsala, Sweden: Association for Computational Linguistics.

Kerremans, Koen. 2012. Translating terminological variation: The case of biodiversity terminology. In L. Zybatow, A. Petrova & M. Ustaszewski (eds.), *Translationswissenschaft: Alte und neue Arten der Translation in Theorie und Praxis* (Forum Translationswissenschaft 16), 71–77. Frankfurt am Main: Peter Lang Verlag.

L'Homme, Marie-Claude. 2003. Capturing the lexical structure in special subject fields with verbs and verbal derivatives: A model for specialized lexicography. *International Journal of Lexicography* 16. 403–422. DOI:doi:10.1093/ijl/16.4.403

Peirsman, Yves, Kris Heylen & Dirk Speelman. 2008. Putting things in order. First and second order context models for the calculation of semantic similarity. In *Proceedings of the 9th Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, 907–916. Lyon.

Resche, Catherine. 2000. Equivocal economic terms or terminology revisited. *Meta* 45(1). 158–173.

Rogers, Margaret. 2007. Terminological equivalence in technical translation: A problematic concept? *Synaps : fagspråk, kommunikasjon, kulturkunnskap* 20. 13–25.

Rychlý, Pavel & Adam Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 41–44. Prague, Czech Republic: Association for Computational Linguistics.

Shimizu, Nobuyuki, Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama & Hiroshi Nakagawa. 2008. Metric learning for synonym acquisition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 793–800. Manchester, UK: Coling 2008 Organizing Committee.

Tanskanen, Sanna-Kaisa. 2006. *Collaborating Towards Coherence: Lexical Cohesion in English Discourse*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Temmerman, Rita. 1997. Questioning the univocity ideal. The difference between socio-cognitive terminology and traditional terminology. *Hermes* 18. 51–90.

Temmerman, Rita. 2000. *Towards new ways of terminology description: The sociocognitive-approach*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Weeds, Julie & Daniel Marcu. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics* 31(4). 439–475.