AJAL

Argentinian Journal of Applied Linguistics

FAAPI

AJAL Argentinian Journal of Applied Linguistics

Contents

Book reviews

Editorial introductions

Introduction to the Special Issue (Part 1) Language and technology

Language and technology have walked hand in hand for a long time. Maybe the first technology that was incorporated to language was the paint in the first caves where signs of distinctive language signs can be found. The history of technology went on with the different methods to be used to write until the humanity reached the press. This, obviously, represents a significant step ahead in the recording of the language. It will not be until the beginning that Henry Mill achieves its first commercial type of typewriter, at the end of the 19th century when we find another significant step with the invention of the phonograph. Many readers will remember the extensive repetitive phonetic exercises of Eliza Doolittle as inducted by Higgins in *Pigmalion*. The 20th century served to facilitate the use of different writing recording systems especially after the arrival of the personal cassette player that has served to many of us to learn languages at home or the walkie-talkie as a ubiquitous device. Since the 1980s the extensive use of computers first and other mobile devices later has helped as a means to communicate, analyze and interact with the language in many different ways that only 50 years ago would have been hard to imagine.

This special volume presents six different articles that deal with different issues of the relation between language and technology. The first, *"Profiling Specialized Web Corpus Qualities: A Progress Report on «Domainhood»"*, by Santini, Strandqvist and Jönsson presents several case studies on specialized or domain-specific web corpora. The paper evidences that it is possible to profile the domainhood quality of a corpus. The second paper entitled *"Extraction of terms for the construction of semantic frames for named bays"* by Rojas-García and Faber is centered on the description of a terminological knowledge that permits the geographic contextualization of linguistic data through a semi-automatic method of extracting terms based on quantitative procedures from a specialized corpus. The results showed that their procedure described the semantic frames of named bays based on Frame-based Terminology. The third paper authorized by Esther de Boe, *"Remote Healthcare Interpreting: a methodology to investigate quality"*, addresses a topic of communication in healthcare that has increased its presence in translations and interpretation in the last fifteen years. The current research and this paper show the increasing importance of technology in assisting this specific type of communication and de Boe's paper provides a

methodological framework in comparing telephone and video interpreting on one side, and face-to-face on the other. The author considers that while the messages through the three channels can be similar, changes in the interactional dynamics had an impact in the communication. A fourth paper that winks at technology is Cenni's *"Multilingualism 2.0: language policies and the use of online translation tools on global platforms"* which considers that the web 2.0 has had a significant role in the interaction of many languages in online collaborative communication. Her study aims to present a number of different topics on multilingualism that have arisen with the advent of the web 2.0 in five platforms such Wikipedia, Facebook, Instagram, Booking.com and TripAdvisor especially the challenge of using a large number of languages to get the information through to its users. In the fifth article, *"Validating multilingual hybrid automatic term extraction for search engine optimization: the use case of EBM-GUIDELINES*", Rigouts et al. make an introspection into the tools that automatically extract terms and their own equivalents in different languages and specifically looks at their use in a specific case from a medical website. In the final paper, *"High-density knowledge rich contexts"*, León-Araúz and Reimerink show how Knowledge Rich Contexts (KRCs) combines items of knowledge supports the job of terminographers and the acquisition of knowledge by the users.

All in all, this interesting volume addresses topics of different nature. But all of them show the increasing importance of technology in terms of language analysis and linguistic knowledge. There is little question that the author will find these articles very appealing and will serve for further studies in the same topics. Having done an extensive selection where only about forty percent of the papers submitted were accepted by the first screening and two more after. Reviewers had to go through deep revisions and it is our hope that this final version will satisfy those more demanding readers, and will be very appealing for the selected readership of the *Argentinian Journal of Applied Linguistics*.

Cristina Calle Martínez – Universidad Complutense, Spain
Beatriz Sedano Cuevas – National Distance University UNED, Spain
Jesús García Laborda – Universidad de Alcalá, Spain

Editorial: Introduction to our regular call for contributions

Readers may have noticed that this edition of our journal is richer than previous issues because of the combination of a special issue and the regular articles we accept as part of our ordinary call for contributions.

The regular issue features three articles by educators from Indonesia and Argentina, and five book reviews. Regarding the book reviews, we should highlight that the reviews authored by Paniagua, Alderete, Ferrario and Lescano are part of an interinstitutional project between higher education institutions in Salta and Chubut and that the authors were student-teachers at the time of writing and submitting their contributions.

Finally, we would like to thank the reviewers who contributed with their expertise and generosity in 2018:

Débora Bagüez

Eugenia Carrión Cantón

Hyeyoon Cho

Christina Gkonou

Bushra Ahmed Khurram

Mario López Barrios

Celeste Luna

María Inés Miller

Carolina Orgnero

Yecid Ortega

Gabriela Tavella

Patricia Weller

Darío Luis Banegas and María Susana Ibáñez

Cover photo from https://voxy.com/blog/2019/02/the-value-of-videos-in-language-learning/

Profiling specialized web corpus qualities: A progress report on "Domainhood"

Marina Santini[*] - RISE Research Institutes of Sweden
Wiktor Strandqvist - RISE Research Institutes of Sweden & Linköping University
Arne Jönsson - RISE Research Institutes of Sweden & Linköping University

This article is an extended version of the following paper:
Strandqvist W., Santini M., Lind L. and Jönsson A. (2018). Towards a Quality Assessment of Web Corpora for Language Technology Applications. In: Read T., Montaner S. and Sedano B. (2018). *Technological Innovation for Specialized Linguistic Domains Languages for Digital Lives and Cultures Proceedings of TISLID'18*. Editions universitaires europeenne.

ABSTRACT

In this article we describe ways to profile the domain specificity, a.k.a. domainhood, of specialized web corpora in English and in Swedish. Several studies have been carried out to measure the "qualities" of general-purpose web corpora. On the contrary, less attention has been paid to the evaluation of specialized or domain-specific web corpora. To fill this gap, in this article we present case studies where we explore the effectiveness of several statistical measures – i.e. rank correlation coefficients (Kendall and Spearman), Kullback–Leibler divergence, log-likelihood and burstiness - to assess domainhood. Our findings indicate that it is possible to profile the domainhood quality of a corpus. However, further research is needed to generalize on the results.
Keywords: corpus evaluation; term extraction; log- likelihood; rank correlation; Kullback-Leibler divergence.

RESUMEN

En este artículo describimos formas de trazar la especificidad del dominio ("domainhood") de los corpus de webs especializados en inglés y en sueco. Muchos estudios se han llevado a cabo para medir las "cualidades" de los corpus de webs de carácter general. Sin embargo, se ha prestado menos atención a la evaluación de corpus de web especializados o de dominios específicos. Para llenar este vacío, en este artículo presentamos estudios de caso donde exploramos la efectividad de diferentes medidas estadísticas, a saber, coeficientes de correlación de rango (Kendall and Spearman), divergencia Kullback–Leibler, probabilidad de registro y *burstiness* – para evaluar la especificidad del dominio. Nuestros resultados indican que es posible perfilar la calidad de dominio de un corpus. Sin embargo, es necesaria una mayor investigación para generalizar en los resultados.
Palabras clave: evaluación de corpus; extracción de términos; probabilidad de registro correlación de rango; divergencia Kullback-Leibler.

_____
* Corresponding author; e-mail: marinasantini.ms@gmail.com, marina.santini@ri.se

WEB CORPORA ARE text collections made of documents that have been retrieved and downloaded from the web. While texts in traditional corpora are hand-picked from several media and agreed upon by a number of experts, web corpora are built with documents available on the web at the time of corpus bootstapping. Traditional corpora are carefully curated and annotated to preserve the original traits of the selected texts, while web corpora can be noisy in several respects, e.g. they might contain damaged characters, problematic symbols, inconsistent punctuation or ungrammatical texts. In short, traditional corpora and web corpora represent different approaches to corpus construction and use.

Traditional corpora are a trove of hand-crafted qualities. However, the added value of web corpora is in their malleability. Similar to traditional corpora, web corpora can be general-purpose or specialized (Barbaresi, 2015) and may serve different purposes, such as linguistic studies (e.g., Schäfer & Bildhauer, 2013; Biemann et al., 2007; Lüdeling et al. 2007) and professional uses (Goldhahn et al., 2012; Baroni et al., 2006). The unique and unprecedented potential of web corpora is that they can promptly and inexpensively account for virtually any domain, topic, genre, register, sublanguage, style and emotional connotation, since the web itself is a gold mine of linguistic and textual varieties. In particular, domain-specific web corpora are widely used in several linguistic disciplines (e.g. in translation studies and lexicography) and they are an important building block of language technology applications (e.g. machine translation, terminology extraction and lexicon induction). Both in linguistics and in language technology, the reliability of the results may depend on the domain representativeness of the web corpus itself.

While bootstrapping a web corpus is common practice (many tools exist, either based on crawling or on search engine queries), the validation of web corpora is still a grey area. With the investigations described in this article, we would like to contribute to the discussion by adding a new perspective to web corpus evaluation. Normally, corpora can be assessed according to several "qualities", for instance corpus balance (in terms of domain, genre, style, register etc.), corpus representativeness (with respect to a purpose), and the like. In this complex scenario, we single out one quality, namely domain specificity, a.k.a. domainhood. Domainhood (Santini et al., 2018) refers to the domain representativeness of a corpus. Here "domain" is defined as the "subject field" or "area" in which a web document is used. For instance, a high frequency of medical terms is a sign that a corpus is a specialized medical corpus. We are aware that domain-specific web corpora might have a different domain granularity, and this is an additional factor to be taken into account.

In this article, we present case studies where we explore the effectiveness of several statistical measures – i.e. rank correlation coefficients (Kendall and Spearman), Kullback–Leibler (KL) divergence, log-likelihood and burstiness - to assess domainhood. The long-term goal is to find suitable metrics that would help assess whether one corpus is more domain-specific than another corpus. This information would speed up any post-editing of specialized web corpora by reducing manual intervention. In this article we empirically

investigate these issues and present two experiments, the first one based on English corpora, and the second one hinged upon Swedish corpora.

The article is organized as follows: in Section 2 ("Related Work"), we discuss previous research; Section 3 ("Experiment 1: Building and Profiling Domain-Specific Web Corpora in English") presents the eCare Term Extractor and the profiling of two medical web corpora in English with similar domain granularity and a similar corpus size; in Section 4 ("Experiment 2: Building and Profiling Domain-Specific Web Corpora in Swedish"), we apply burstiness to pin down the domainhood of two medical web corpora in Swedish that have different domain granularity and a different size; finally in Section 5 ("Conclusion and Future Work"), we draw conclusions and outline future work.

Related Work

When we talk about web corpora, it seems more appropriate to talk about "qualities" rather than a single "quality". Several approaches have been proposed to capture the "qualities" of web corpora (e.g. see Oakes, 2008; Schäfer et al., 2013). However, no standard metrics have been agreed upon for the automatic quantitative assessment of the different "qualities" of web corpora. "Qualities" can be defined as dimensions of variation. Domain, genre, style, register, medium, etc. are well-known dimensions of language variation. In this study, we focus on the dimension of "domain", i.e. the "subject field" in which a web document is used. Our aim is somewhat similar to the one expressed in Wong et al. (2011), where the authors propose a technique, called SPARTAN, for constructing specialized corpora from the web. Our approach is different though, because in order to assess the domainhood quality, we rely on measures that are well-established and easy to replicate. Since in this article we describe comparative experiments based on rank correlation (Kendall and Spearman), KL divergence, log-likelihood and burstiness, in this section we provide a short overview of studies where these measures were used.

The importance of a quantitative evaluation of corpora has been stressed for a long time. In his seminal article, Kilgarriff (2001) motivates his review of approaches to corpus comparison by asking two crucial questions: "how similar are two corpora?" and "in what ways do two corpora differ?". He presents comparative experiments based on several corpora and on several statistical measures. Rayson and Garside (2000) show that log-likelihood can be safely used as a "quick way to find the differences between corpora" and that it is more robust than other measures because it is insensitive to corpus size. Gries (2013) suggests using a Kendall Tau correlation coefficient to determine whether the observed patterns of two corpora show significant correlations. Ciaramita and Baroni (2006) propose using KL divergence to assess the "randomness" or "unbiasedness" of general-purpose corpora. They compare domain-specific sub parts of the British National Corpus (BNC) against the whole corpus and show that KL divergence can reliably indicate the difference between general purpose corpora (random and unbiased) and domain-specific

corpora (biased). Burstiness has been used in information retrieval and in terminology extraction (Church and Gale, 1995; Katz, 1996), and more recently for corpus evaluation (Sharoff, 2017). Burstiness is a measure that can be utilized for inducing specialized lexicon that is not evenly distributed in a corpus but appears "in bursts". Burstiness indicates "how peaked a word's usage is over a particular corpus of documents" (Pierrehumbert, 2012). More specifically, "bursty words are topical words that tend to appear frequently in documents when some topic is discussed, but do not appear frequently across all documents in a collection" (Irvine and Callison-Burch, 2017). While bursty words are feared and filtered out when assessing general-purpose corpora (Sharoff, 2017), we think that they could give a good indication of domain specificity, and for this reason we include burstiness in our experiments.

Experiment 1: Building and Profiling Domain-Specific Web Corpora in English

In the first experiment, we propose a two-step approach. In the first step, we build a term extraction that can automatically identify term candidates in project-specific personas and use cases/scenarios. Personas and use cases/scenarios are narratives that describe a "system's behavior under various conditions as the system responds to requests from stakeholders" (Cockburn, 2000). This type of texts are nowadays normally included in many language technology projects. Personas and use cases/scenarios are relatively short texts - only a few dozen pages − normally based on numerous interviews and observations of real situations and written by domain experts who know how to correctly use terms in their own domain. For this reason, we argue that they are a convenient textual resource to automatically extract term seeds to bootstrap domain-specific web corpora, thus overriding the tedious and somehow arbitrary process normally required to collect term seeds. In our study, we focus on the medical terms that occur in personas and use cases/scenarios written in English for *E-care@home*, a multi-disciplinary project that investigates how to ensure medical care at home for the elderly. We complete this step with the evaluation of the term extractor against a gold standard made of the SNOMED CT terms manually selected by a domain expert from the *E-care@home* personas and use cases/scenarios. SNOMED CT is the largest existing resource of medical terminology. The challenge of this step is to create a "good enough" term extractor based on a relatively small textual resource, a task that is still under-investigated since most of existing term extractors are based on large corpora (e.g. see Nazarenko and Zargayouna, 2009).

In the second step, we use the term seeds extracted in the previous step to *bootcat* a medical web corpus and evaluate its quality. The term "bootcat" means bootstrapping specialized corpora from the web using BootCaT, a tool that was introduced by Baroni and Bernardini (2004). Leveraging on the web to create specialized corpora is a well-established idea (e.g. Baroni & Bernardini, 2004; Kilgarriff et al. 2010), less so their evaluation. For this reason, in Experiment 1 we analyse and test three corpus profiling measures, namely rank

correlation (Kendall and Spearman), KL divergence and log-likelihood. The challenge of this step is to find an empirical answer to the following question: "*can we assess the domainhood quality of a corpus automatically bootstrapped from the web*?".

E-care Term Extractor

Arguably, the use of personas and use cases/scenarios, when available, is a good starting point to automatize the manual process of term seeds selection. The *E-care* term extractor developed for this purpose includes three main components. The first component (*terminology extractor*) uses a shallow syntactic analysis of the text to extract candidate terms. The second component (*terminology validator*) compares each of the candidate terms and their variations against SNOMED CT (International Edition) to produce candidate terms. The third component (*seed validator*) evaluates the performance of the term extractor.

The *terminology extractor* relies on the Stanford Tagger (Toutanova, Klein, Manning, & Singer, 2003) to assign a part-of-speech (POS) tag to each word in the texts. The tagged text is then searched sequentially with each of the selected syntactic patterns shown in Table 1 (cf. Pazienza, Pennacchiotti, & Zanzotto, 2005).

Patterns
(noun)+
(adjective)(noun)+
(noun)(prep)(noun)+

Table 1. Syntactic patterns used to identify terminology

The *terminology validator* takes the candidate terms produced in the previous steps, and matches them against SNOMED CT. If an exact match is not found, each word is stemmed. The stemmed words are permuted, and each permutation is then matched against SNOMED CT once again, this time using wildcards between the words, to allow for spelling variations. Matches are then ranked by DF/IDF scores (cutoff = 200). In this context, DF stands for *term document frequency* and refers to the frequency of a term in a document divided by the document length (i.e. the total number of words in the document). DF is basically like a TF (*term frequency*[1]) but normalized to the document length in order to avoid any bias towards long documents[2]. IDF stands for *inverse document frequency* and it is based on counting the number of documents in the corpus which contain the term in question (Robertson, 2004). Similar to TF-IDF, DF/IDF is a way to reflect how important a term is in a given document.

The *seed generator* generates three terms (i.e. triples) from the cutoff list when they occurred in the same document.

E-care Term Extractor: Results and Discussion

The *E-care* term extractor performance is summarized in Table 2. The *terminology extractor* has an extraction recall of 81.25% on the development set, which is the subset of documents used to optimize the extraction algorithm. When evaluated against the gold standard made of the SNOMED CT terms manually selected by a domain expert from the E-care@home personas and use cases/scenarios, the *terminology validator* achieves the following performance: Precision = 34.2%, Recall = 71%, F1 = 46.2%. These metrics were calculated according to standard formulas, namely Precision = true positives / (true positives + false positives), Recall = true positives / (true positives + false negatives) [3] and F1 = 2 * ((precision * recall) / (precision + recall))[4].

	Metrics	%
Term candidate extraction	Extraction recall	81
Term validation	Precision	34.2
	Recall	71
	F1	46.2

Table 2. Current performance of E-care term extractor

Interestingly, the moderate performance of the current version of the *E-care* term extractor did not affect detrimentally the domainhood quality of the resulting web corpus, as shown in the next subsections.

Corpus Evaluation Metrics

For corpus evaluation, we use metrics based on word frequency lists, namely rank correlation coefficients (Kendall and Spearman), KL divergence and log-likelihood.

1) Correlation coefficients: *Kendall* correlation coefficient (Tau) and *Spearman* correlation test (Rho) are non-parametric tests. They both measure how similar the order of two ranks is. We used the R function "cor.test()" with method="kendall, spearman" to calculate the tests.

2) Kullback–Leibler (KL) divergence (a.k.a. relative entropy): KL divergence is a measure of the "distance" between two distributions. The KL divergence quantifies how far-off an estimation of a certain distribution is from the true distribution. The KL divergence is non-negative and equal to zero if the two distributions are identical. In our context, the closer the value is to 0, the more similar two corpora are. We used the R package "entropy", function "KL.empirical()" to compute KL divergence.

3) Log-Likelihood (LL-G^2): log-likelihood (Dunning, 1993) has been used for corpus profiling (Rayson and Garside, 2000). The words that have the largest log-likelihood scores

show the most significant word-frequency difference in two corpora. Log-likelihood is not affected by corpus size variation.

For the evaluation, we use three web corpora, namely:

- **ukWaCsample** (872 565 words): a random subset of ukWaC (Ferraresi et al., 2008), a general- purpose web corpus.
- **Gold** (544 677 words): a domain-specific web corpus bootstrapped with ***terms manually selected by a domain expert*** from the *E-care@home* personas and use cases/scenarios.
- **Auto** (492 479 words): a domain-specific web corpus collected with ***automatically extracted term seeds*** from the *E-care@home* personas and use cases/scenarios.

Results and Discussion

In this section, we present and discuss the results of our experiments.

Measuring Rank Correlation. We computed the normalized frequencies of the three corpora (words per million) and ranked them (with ties). The plots of the first 1000 top frequencies of the three corpora are shown in Fig. 1. From the plots, we can see that UkwaCsample has very little in common with both Gold and Auto (boxes 1 and 2), while Gold and Auto (box 3) are similar.

Fig. 1 Plotting 1000 top ranks: (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

When testing the rank correlation (Kendall and Spearman), we observe a statistically significant positive rank correlation between Gold and Auto (see Fig. 2, box 3; Fig. 3, box 3), which means that words in Gold and in Auto tend to have similar ranks. Conversely, the correlation between ukWaCsample and Gold and ukWaCsample and Auto is negative and weak (see Fig. 2, box 1 and box 2; Fig. 3, box 1 and box 2), which essentially means that their ranks follow different distributions.
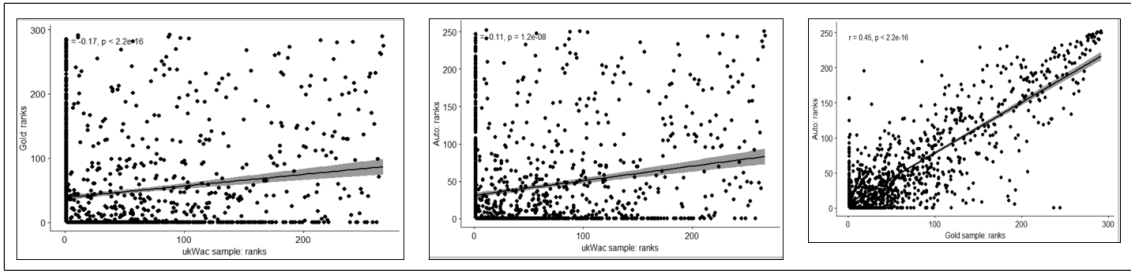
Fig. 2 Kendall Tau: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).
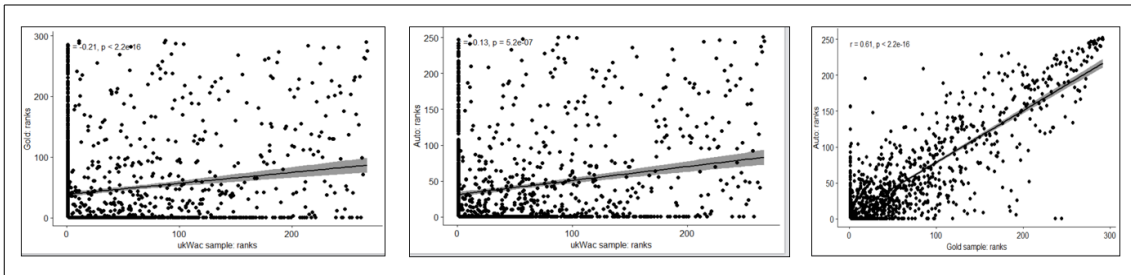
Fig. 3 Spearman Rho: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

KL divergence. Before calculating KL divergence, a smoothing[5] value of 0.01 was been added to the normalized frequencies. Results are shown in Table 3. The scores returned by KL distance for ukWacSample vs Gold (row 1) and ukWacSample vs Auto (row 2) − 7.544118 and 6.519677, respectively − are (unsurprisingly) large and indicate a wide divergence between the general-purpose ukWacSample and the domain-specific Gold and Auto. On the contrary, the KL score of 1.843863 indicates that Gold vs Auto (row 3) are similar to each other.

Corpora	KL scores
ukWacSample vs Gold	7.544118
ukWacSample vs Auto	6.519677
Gold vs Auto	1.843863

Table 3. KL scores

Log-Likelihood (LL-G^2). We computed log-likelihood scores on smoothed word frequencies. The total log-likelihood scores for the three web corpora (top 1000 words) are shown in Table 3. The larger the log-likelihood score of a word, the more different its distribution in two corpora. The large log-likelihood scores for ukWaCsample vs Gold (453 441.6) and for ukWaCsample vs Auto (393 705.9) indicate that these corpora are remarkably dissimilar if

compared to the much smaller log-likelihood score returned for Gold vs Auto (114 694.2), which suggests that Gold and Auto are similar to each other (see Table 4).

Corpora	Total log-likelihood scores
ukWacSample vs Gold	453 441.6
ukWacSample vs Auto	393 705.9
Gold vs Auto	114 694.2

Table 4. Log-likelihood scores of the three corpora

It is also possible to assess the statistical significance of the individual log-likelihood scores. Normally, a log-likelihood score of 3.8415 or higher is significant at the level of $p<0.05$ and a log-likelihood score of 10.8276 is significant at the level of $p<0.001$ (Desagulier, 2017). Fig. 4 shows the breakdown of the top-ranked log-likelihood scores of three corpora. We take 3.8415 ($p < 0.05$) as a threshold and observe that ukWaCsample vs Gold (box 1) differs very much in the use of words such as "patient" or "patients" and "blood", and in ukWaCsample vs Auto (box 2) these words have a similar behavior. Conversely, these words are not in the top list of Gold vs Auto (box 3). Additionally, the log-likelihood scores in box 3 are much smaller in magnitude, which indicates that the difference between words is less pronounced.

patients	6162.61		blood	5825.56		headache	1040.9
blood	5092.01		risk	3847.85		parkinson's	967.5
patient	4120.3		patients	3827.24		valve	680.56
symptoms	3803.51		diabetes	3725.77		aortic	677.53
disease	3654.71		heart	3657.2		milk	535.3
treatment	3326.27		pressure	2868.36		ltot	453.59
risk	3121.56		pain	2867.39		eggs	426.37
heart	2959.02		oxygen	2730.52		administration	420.84
stroke	2733.91		symptoms	2581.23		stenosis	402.16
diabetes	2712.8		patient	2286.48		memory	396.69
			disease	2198.23		online	396.11
			glucose	2071.23			

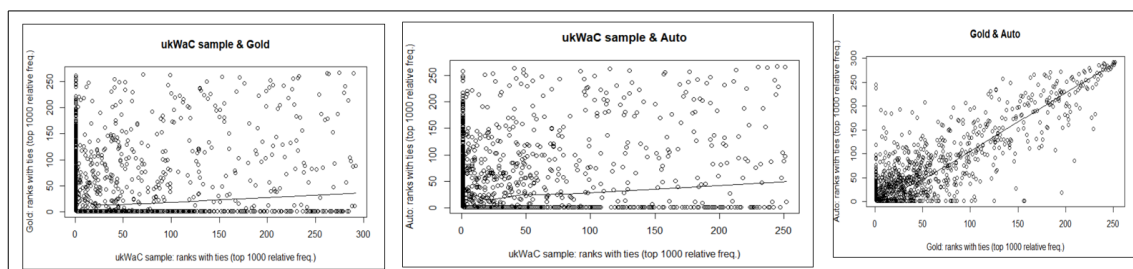Fig. 4. Top-ranked log-likelihood scores (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box 2), and Gold and Auto (box 3).

Experiment 1: Conclusion

We have shown that it is possible to create a reliable term extractor (although the intrinsic evaluation of its performance is moderate) that works well in practice for relatively short texts written by domain experts. When used to bootstrap a web corpus, the automatically extracted term seeds create a corpus whose domain specificity is similar to a corpus bootstrapped with manually selected term seeds. This is an added value because corpus construction can be fully automatized and standardized.

We have seen that well-established measures, such as rank correlation, KL divergence and log-likelihood, do give a coarse but grounded idea of domain specificity. Essentially, they allow for an evaluation of the domainhood quality of web corpus and presumably they could also be used to pre-assess the portability of NLP tools from a domain-specific corpus to another. Similar experiments have also been carried out on Swedish corpora with much the same results (Santini et al., 2018), showing that our approach may become a language-independent standardized step in corpus evaluation practice, if these results will be confirmed by future experiments in other languages.

We can now provide empirical answer to the questions asked above. That is: yes, we can assess the domainhood quality of a corpus automatically bootstrapped from the web. This can be done by using metrics that are well-established and easily replicable, such as rank correlation, KL divergence and log-likelihood. Last but not least, these metrics also help get a coarse but robust indication of topical similarities across corpora.

Experiment 2: Building and Profiling Domain-Specific Web Corpora in Swedish

Since "words are not selected at random" (Kilgarriff, 2005), we assume that the content words included in a corpus represent its content and domain. The corpora that we describe below both belong to the medical domain, but they have been built with slightly different target domains and domain granularity. The target domains are reference lists made of words representative of the domain of interest. As pointed out by Lippincott et al. (2011) "[w]hile variation at a coarser domain level such as between newswire and biomedical text is well-studied and known to affect the portability of NLP systems, there is a need to develop an awareness of subdomain variation when considering the practical use of language processing applications". In this experiment, we investigate whether burstiness can help make sense of subdomain variations or different domain granularities.

Same Domain, Different Granularities

For this investigation, we rely on two web corpora of Swedish texts, namely *eCare_ch_sv_01* and *eCare_uc_sv_02*. Both corpora are components of the eCare web corpus. *eCare_ch_sv_01* is about chronic diseases, while *eCare_uc_sv_02* was built with terminology in English automatically extracted in Experiment 1, and then translated in via SNOMED CT.

eCare_ch_sv_01 was built using 155 terms listed in SNOMED CT (Swedish edition) indicating chronic diseases. The 155 term seeds were selected from a much longer list of chronic diseases compiled by a domain expert and they represent a restricted and fine-grained domain (Santini et al., 2017). The size of this corpus is approx. 700 000 words. This corpus was used in the experiments presented in Santini et al. (2018).

The size of *eCare_uc_sv_02* is approx. 7 million words (6 942 193 tokens). *eCare_uc_sv_02* is, thus, about 10 times larger than *eCare_ch_sv_01*.

Both web corpora are supposed to represent the domain of chronic diseases but with different domain granularities and different corpus sizes. We assume that the domain granularity is more fine-grained in *eCare_ch_sv_01* and coarser in *eCare_uc_sv_02* because of the way the corpora have been bootstrapped. In this experiment, "fine-grained domain" means a very specialized domain where the seeds to bootstrap the corpus are specialized medical terms, e.g. "artrit" (en: arthritis). Conversely, "coarse-domain" refers to a corpus that has been bootstrapped both with specialized medical terms and polysemous words that are often related with diseases, e.g. "dos" (en: dosage) or "akut" (en: acute). The domain granularity is implicitly incorporated in the gold standards, as explained below.

Corpus Seeds and Gold Standards

What is the best way to represent a target domain? This question is complex and arguably the ideal solution depends on the purpose of an application. Here we take a basic approach and represent the target domains as reference lists – the gold standards - that contain the term seeds used to bootstrap the corpora. It makes sense to use domain-specific terms both for bootstrapping a web corpus and for evaluating its domainhood because the terms used as seeds (source terms) should be found in non-trivial proportions to be sure that the corpus is representative of the domain of interest. Here we present two different approaches to gold standard construction. The gold standard used to profile and evaluate *eCare_ch_sv_01* is made only of specialized medical terms, while the gold standard automatically extracted from use cases contains also polysemous words, such as "attack" (en: attack), "extrem" (en: extreme), "fet" (en: fat). The gold standards contain tokenized term seeds, without duplicates. This means that terms like "kronisk anemi" (en: chronic anemia) and "kronisk artrit" (en: chronic arthritis), in the gold standard are represented by three entries, namely "kronisk", "anemi" and "artrit". Both these lists and the top-ranked bursty words were stemmed, stopwords and numbers were removed using the R package Quanteda, without applying any customization to the stoplist and to the stemmer.

The two web corpora are then evaluated against two gold standards. More specifically, *gold_eCare_ch_sv_01* represents the target domain of *eCare_ch_sv_01* and contains 164 unigrams, while the target domain of *eCare_uc_sv_02* is represented by *gold_eCare_uc_sv_02* that contains 248 unigrams.

Burstiness

Several burstiness formulas exist. Here we use the formula from Church and Gale (1995), including the modification proposed by Irvine and Callison-Burch (2017) (i.e. the use of relative frequencies rather than absolute frequencies), namely:

$$B_w = \frac{\sum_{d_i \in D} rf_{w_{d_i}}}{df_w}$$

where *rf* refers to the relative frequency of word *w* in a document, and *df* is the number of documents in which the word *w* appears. Relative frequencies are raw frequencies normalized by document length. In other words, burstiness is given by the sum of the all the relative frequencies of word *w* in the documents of the corpus divided by the number of documents containing the word. Burstiness is essentially the mean of a word in a corpus normalized by the number of documents where the word appears, and it ignores the documents where the word does not appear (Church and Gale, 1995; Katz, 1996). Burstiness differs from measures like TF (term frequency), which denotes the number of times that term *t* occurs in document *d*, or TF-IDF, a weight where TF is normalized by IDF (inverse document frequency)[6]. If compared with profiling measures such as log-likelihood, burstiness is a "self-contained" measure, because it does not need a reference corpus to be calculated, and the top-ranked bursty words can be easily matched against a gold standard representing the target domain.

Assessment of Bursty Words against Gold Standards

Burstiness was calculated separately for *eCare_ch_sv_01* and for *eCare_uc_sv_02*. For each corpus, we sorted the burstiness values by decreasing order and we took the top 2105 bursty words for *eCare_ch_sv_01* (Santini et al., 2018) and the top 21028 bursty words for *eCare_uc_sv_02* (since *eCare_uc_sv_02* is about 10 times larger than *eCare_ch_sv_01*) and matched them against the two gold standards that were described above. We used several metrics to assess the results, namely: intersection, percentage, precision@, Jaccard and Dice coefficients. For precision@ we use two cut-off points, i.e. 2105 for *eCare_ch_sv_01* and 21028 for *eCare_uc_sv_02*.

	Inter	**%**	**Precision@**	**Jaccard**	**Dice**
ch_sv_01	93	58.1%	0.0359	0.0427	0.0819
uc_sv_02	183	73.7%	0.0111	0.0086	0.0172

Table 5. Assessment of bursty words against gold standards.

Results are shown in Table 5, which reports the intersection between the top-ranked scores and the gold standard (col.2), percentage (col. 3), precision@ (col. 4), Jaccard coefficient (col.5), and Dice coefficient (col. 6). The size of the intersection and the percentage give an intuitive understanding of the overlap between the top-ranked bursty words and the target domains stored in the gold standards. The intersections amount to 58.1% for *eCare_ch_sv_01* and 73.6% for *eCare_uc_sv_02*. We think these figures are promising because when we measured the bursty words extracted from the Swedish National Corpus (called Stockholm-Umeå Corpus or SUC), the intersection with *gold_eCare_ch_sv_01* amounted to one occurrence (Santini et al., 2018), as shown in Table 6.

	Intersection	Jaccard	Dice	Precision@2105
SUC	1	0.000440	0.00088	0.00001

Table 6. Intersection between SUC bursty words and *gold_eCare_ch_sv_01*.

It is also worth noting that burstiness seems to be robust to corpus size variation since we observe that the number of domain-specific words identified increases with the size of the corpus rather than dropping. Apparently, the values of precision@ and those of Dice and Jaccard coefficients do not make justice to the magnitude of the overlap since their calculation takes into account the number of unmatched items, which in our case are many because the gold standards are much shorter than the lists of top-ranked bursty words.

Discussion

Results show that burstiness and the extent to which words with a higher burstiness overlap with gold standards (i.e. reference lists comprising domain-specific vocabulary) can be used to profile and quantify the domain specificity of a (web) corpus. As stated earlier, the burstiness of a word indicates the extent to which its frequency is unevenly distributed across documents within a specialized web corpus. This characterization fits very well the web corpora used in these experiments where domain-specific medical terms appear only in some documents and are not evenly distributed in all the documents of the corpus. We find these results auspicious because burstiness has the potential to "discover" and bring to the surface words that are important and domain- specific, but that could be missed out by other metrics, like log-likelihood, because they are distributed unevenly across a corpus (see also results in Santini et al., 2018). In a situation like this one, also a measure like perplexity, an evaluation metric that is often used to evaluate language models and that is also employed to assess domain adaptation in NLP tasks, could give misleading results, because it can be biased by the number of "unpredictable" bursty words.

In Experiment 2, many bursty words match the gold standards. This is encouraging because burstiness seems to capture the way in which content is distributed in this kind of web corpora. We observe that an intersection of 93 words out of the 160 unigrams listed in

gold_eCare_ch_sv_01 (58.1%) indicates that about 8% of the 2015 top-ranked bursty words belong to the fine-grained domain of 155 SNOMED CT chronic diseases. An intersection of 183 words out to the 248 unigrams listed in *gold_eCare_uc_sv_02* (73.7%) indicates that about 1.2% of the 21028 top-ranked bursty words belong to the coarse-grained domain extracted from eCare use cases (see Table 7).

	eCare_ch_sv_01	*eCare_uc_sv_02*
Coprus size (words)	700 000	7 000 000

Corpus Seeds	155 chronic diseases. Ex:
	`[...]`
	`lungemfysem`
	`mycetom`
	`ozena`
	`polyserosit`
	`postkardiotomisyndrom`
	`Swimmingpooldermatit`
	`trumhinneatelektas`
	`adhesiv mediaotit`
	`aktinomykotiskt mycetom`
	`[...]`
		160 triplets extracted from uses cases. Ex:
		`[...]`
		`tremor "parkinsons sjukdom" styrka`
		`skakningar "parkinsons sjukdom" styrka`
		`urinanalys "fynd som rör viktförlust" lever`
		`koma "akut exacerbation av kroniskt obstruktiv luftvägssjukdom" hjärtfrekvens`
		`"angina pectoris" angina balanserad`
		`uppföljningsstatus insulin "skyddat boende"`
		`långtidssyrgasbehandling medicin sarkoidos`
		`blåsa "diabetes mellitus typ 2" riskbedömning`
		`pillerask "mätning av fysiologiska parametrar" sjukgymnast`
		`[...]`

Gold Standard (GS)	160 unigrams	248 unigrams
Top-ranked bursty words (BW)	2105 unigrams	21028 unigrams
Intersection: GS & BW	93 words out of 160 (58.1%)	183 words out of 248 (73.7%)
% of BW in GS	approx. 8% $[2105:100=160:x]$	approx. 1.2% $[21048:100=248:x]$
% of BW in Intersection	approx. 4% $[2105:100=93:x]$	approx. 0.9% $[21248:100=183:x]$

Table 7. Summary table

At this stage of research, we do not make any assumption about the minimum size of intersection that would account for a certain domain granularity, since we need further investigations to find a more principled approach to assess the relation between the size of the corpus, the length of the gold standards, and the cut-off points.

Open Issues

Research on the quantification of domain granularity of corpora bootstrapped from the web is still at the outset and several issues need to be further discussed and investigated.

Domain granularity: we put forwards two working definitions, namely "fine-grained domain" means bootstrapped with specialized medical terms, and "coarse-grained domain" means bootstrapped with both specialized medical terms and more general words.

Evaluation: the quantification using the intersection and percentage is more intuitive than precision@, Jaccard and Dice coefficients. However, further experimentation is needed to establish a balanced and principled relation between the size of the corpus, the length of the gold standards, and the cut-off points.

Cut-off points: the decision about the cut-off points was based on a rule of thumb, but in the future we would rather find more theoretically-grounded threshold settings, for example, the statistical significance of the burstiness scores.

Gold standards: the design of the gold standards is exploratory rather than principled. Discussion with domain experts is ongoing.

Last but not least, here we focus on lexical items because words are easy to pre-process. However, domain specificity certainly includes other aspects, such as special syntactic constructs, stance or sublanguage variations.

Experiment 2: Conclusion

In this experiment, we explored whether burstiness is a suitable measure to profile and quantify domain specificity both for small and large specialized web corpora with different domain granularities. Results show that burstiness can provide an indication of domainhood. We find these results promising because burstiness has the potential to discover terms that are domain-specific but evenly distributed, in a corpus and could easily be ignored by other statistical measures.

However, some open issues need to be further investigated, such as the need for more appropriate evaluation metrics, the quest of less empirical cut-off points, and a more principled design of the gold standards.

Conclusion and Future Work

In this article we have presented two experiments where we have explored measures to assess domainhood in web corpora. Domainhood indicates the degree of domain specificity of a specialized corpus.

In Experiment 1 (English corpora), we have profiled two specialized medical web corpora against a general-purpose web corpus using well-established measures, such as rank correlation, KL divergence and log-likelihood. These measures provide an indicative idea of domain specificity and allow us to assess whether a corpus bootstrapped from the web is satisfactorily domain-specific or whether it needs some amends before being used for linguistic studies or LT applications.

In Experiment 2 (Swedish corpora), we have used burstiness to identify domain-specific terms and lexicon that could give hints about the domain granularity of a corpus. Results show that burstiness can give an indication of the domainhood of a web corpus about diseases, since it helps ferret out terms that are domain-specific, but that could be ignored because of their uneven distribution.

The statistical measures that have been tried out in the experiments seem to be language-independent, since they give similar results for English and Swedish.

We are currently planning several follow-up studies that include comparative experiments between burstiness, perplexity, TF, TF-IDF and topic models on several (web) corpora characterized by different word frequency distributions (e.g. poisson mixtures).

TF can be simply the raw count of a term in a document or it can be the results of different types of normalization. See Wikipedia article about TF-IDF <https://en.wikipedia.org/wiki/Tf-idf>. Retrieved 25 March 2019.

For a different definition of DT, see <https://nlp.stanford.edu/IR-book/html/htmledition/inverse-document-frequency-1.html>. Retrieved 25 March 2019.

See Wikipedia article about Precision and Recall <https://en.wikipedia.org/wiki/Precision_and_recall>. Retrieved 25 March 2019.

See Wikipedia article about F1 <https://en.wikipedia.org/wiki/F1_score>. Retrieved 25 March 2019.

"Probability smoothing is a language modeling technique that assigns some non-zero probability to events that were unseen in the training data. This has the effect that the probability mass is divided over more events, hence the probability distribution becomes more smooth.". (Hiemstra, 2009)

See Wikipedia article about TF-IDF <https://en.wikipedia.org/wiki/Tf-idf>. Retrieved 25 March 2019.

References

Barbaresi, A. (2015). Ad hoc and general-purpose corpus construction from web sources. Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Linguistics. École Normale Supérieure de Lyon (Université de Lyon), France.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. LREC 2004 - Fourth International Conference On Language Resources And Evaluation.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In Proceedings of EAMT (pp. 247-252).

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., ... & Zesch, T. (2013). Scalable Construction of High-Quality Web Corpora. JLCL, 28(2), 23-59.

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. Natural Language Engineering, 1(2), 163-190.

Ciaramita, M., & Baroni, M. (2006). A Figure of Merit for the Evaluation of Web-Corpus Randomness. EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics.

Cockburn, A. (2000). *Writing effective use cases, The crystal collection for software professionals*. Addison-Wesley Professional Reading. (24th printing, 2012).

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R*. Springer.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, vol. 19, no. 1.

Ferraresi, A., Zanchetta, E., Bernardini, S. & Baroni, M. (2008). Introducing and evaluating ukWaC, a very large Web-derived corpus of English. Proceedings of the 4th Web as Corpus Workshop (WAC-4) "Can we beat Google?".

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In LREC (Vol. 29, pp. 31-43).

Gries, S. Th. (2013). Elementary statistical testing with R. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, Cambridge University Press.

Hiemstra D. (2009) Probability Smoothing. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Irvine, A., Callison-Burch, C.: A comprehensive analysis of bilingual lexicon induction. Computational Linguistics 43(2), 273{310 (2017)

Katz, S.M. (1996). Distribution of content words and phrases in text and language modelling.*Natural Language Engineering, 2*(1).

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, Vol.6(1).

Kilgarriff, A., Reddy S., Pomikálek J. and PVS A. (2010). A corpus factory for many languages.Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. *Language and Computers, 59*, 7.

Nazarenko, A., & Zargayouna, H. (2009). Evaluating term extraction. International Conference Recent Advances in Natural Language Processing (RANLP'09).

Oakes, M. P. (2008). Statistical measures for corpus profiling. In Proceedings of the Open University Workshop on Corpus Profiling.

Pazienza, M., Pennacchiotti, M., & Zanzotto, F. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. Springer.

Pierrehumbert, J.B.: Burstiness of verbs and derived nouns. In: Shall We Play the Festschrift Game?, pp. 99-115. Springer (2012)

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora.*

Santini M., Jönsson A., Nyström M. and Alirezai M. (2017). A Web Corpus for eCare: Collection, Lay Annotation and Learning - First Results. Workshop LTA'17 (Language Technology Applications 2017) co-located with FedCSIS 2017, Prague. In M. Ganzha and L. Maciaszek and M. Paprzycki (eds), Position Papers of the 2017 Federated Conference on Computer Science and Information Systems, Proceedings, Vol. 12. pp. 71-78.

Santini M., Strandqvist W., Nyström M., Alirezai M. & Jönsson A. (2018). "Can we Quantify Domainhood? Exploring Measures to Assess Domain specificity in Web Corpora". TIR 2018 - 15th International Workshop on Technologies for Information Retrieval.

Schäfer, R., & Bildhauer, F. (2013). Web corpus construction. *Synthesis Lectures on Human Language Technologies, 6*(4), 1-145.

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2013). The good, the bad, and the hazy: Design decisions in web corpus construction. Proceedings of the 8th Web as Corpus Workshop.

Sharoff, S. (2017). Know thy corpus! Exploring frequency distributions in large corpora. In Diab, M., Villavicencio, A. (eds.) Essays in Honor of Adam Kilgarriff. Text, Speech and Language Technology Series, Springer (2017).

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol.1.

Wong W., Liu W., and Bennamoun M. (2011). Constructing specialized corpora through analysing domain representativeness of websites. *Language resources and evaluation*, Vol.45(2).

Acknowledgements

Extraction of terms for the construction of semantic frames for named bays

Juan Rojas-Garcia[*] - University of Granada (Spain)
Pamela Faber - University of Granada (Spain)

ABSTRACT

EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data. For the geographic contextualization of LANDFORM concepts, this paper presents a semi-automatic method of extracting terms associated with named bays (i.e., *Greenwich Bay*).

Terms were extracted from a specialized corpus, where named bays were automatically identified. Statistical procedures were applied for selecting both terms and bays in distributional semantic models to construct the conceptual structures underlying the usage of named bays. The bays sharing associated terms were also clustered and represented in the same conceptual network.

The results showed that the method successfully described the semantic frames of named bays with explanatory adequacy, according to the premises of Frame-based Terminology.

Keywords: Named bay, Conceptual information extraction, Geographical contextualization, Text mining, Frame-based Terminology.

RESUMEN

EcoLexicon (http://ecolexicon.ugr.es) es una base de conocimiento terminológica sobre ciencias medioambientales, cuyo diseño permite la contextualización geográfica de conceptos de la categoría ACCIDENTE GEOGRÁFICO. Para tal fin, este artículo presenta un método semiautomático para extraer términos asociados con bahías con nombre propio (e.gr., *Bahía de Pensacola*).

Los términos se extrajeron de un corpus especializado, donde las designaciones de bahías se identificaron automáticamente. Se aplicaron procedimientos estadísticos para seleccionar bahías y términos, que se proyectaron en espacios semánticos vectoriales, y se emplearon para construir las estructuras conceptuales que subyacían en el uso de la bahías.

Los resultados muestran que el método es apropiado para describir los marcos semánticos que evocan las bahías, según las premisas de la Terminología basada en Marcos.

Palabras clave: Bahía con nombre propio, Extracción de información conceptual, Contextualización geográfica, Minería de textos, Terminología basada en Marcos.

---

* Corresponding author, e-mail: juanrojas@ugr.es

THE ELECTRONIC RESOURCE EcoLexicon is a multilingual, terminological knowledge base on environmental science (http://ecolexicon.ugr.es) that is the practical application of Frame-based Terminology (Faber, 2012). Since most concepts designated by environmental terms are multidimensional (Faber, 2011), the flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas (León-Araúz, Reimerink & Faber, 2013). However, the geographic contextualization of LANDFORM concepts depends on knowing which terms are semantically related to each landform, and how these terms are related to each other.

This paper presents a semi-automatic method of extracting terms associated with named bays (i.e., *Escambia Bay*) as a type of landform from a corpus of English Coastal Engineering texts. The aim is to represent that knowledge in a semantic network in EcoLexicon according to the theoretical premises of Frame-based Terminology.

The rest of this paper is organized as follows. Section 2 provides motivations for the research, and background on distributional semantic models and clustering techniques. Section 3 explains the materials and methods applied in this study, namely, the automatic identification of named bays, the selection procedures for terms and bays in distributional semantic models, and the clustering technique for bays sharing associated terms. Section 4 shows the results obtained. Finally, Section 5 discusses the results and presents the conclusions derived from this work as well as plans for future research.

Background and Literature Review

Motivations for the Research

Despite the fact that named landforms, among other named entities, are frequently found in specialized texts on environment, their representation and inclusion in knowledge resources has received little research attention, as evidenced by the lack of named landforms in terminological resources for the environment such as DicoEnviro[2], GEMET[3] or FAO Term Portal[4]. In contrast, AGROVOC[5] contains basically a list of named landforms with hyponymic information, whereas ENVO[6] provides descriptions of the named landforms with only geographic details, and minimal semantic information consisting of the relation *located_in* (and *tributary_of* in the case of named rivers and bays).

So far, knowledge resources have limited themselves to representing concepts such as BAY, RIVER or BEACH, on the assumption that the concepts linked to each of them are applicable, respectively, to all named bays, rivers and beaches in the real world. This issue is evident in the following description of forcing mechanisms acting on suspended sediment concentrations (SSC) in bays and rivers.

According to Moskalski and Torres (2012), temporal variations in the SSC of bays and rivers are the result of a variety of forcing mechanisms. River discharge is a primary controlling factor, as well as tides, meteorological forcing (i.e., wind-wave resuspension,

offshore winds, storm and precipitation), and human activities. Several of these mechanisms tend to act simultaneously. Nonetheless, the specific mix of active mechanisms is different in each bay and river. For example, SSC in San Francisco Bay is controlled by spring-neap tidal variability, winds, freshwater runoff, and longitudinal salinity differences, whereas precipitation and river discharge are the mechanisms in Suisun Bay. In Yangtze River, SSC is controlled by tides and wind forcing, whereas river discharge, tides, circulation, and stratification are the active forcing mechanisms in York River.

Consequently, in a knowledge resource, a list of forcing mechanism concepts semantically linked to BAY and RIVER concepts would not represent the knowledge really transmitted in specialized texts. To cope with this type of situation, terminological knowledge bases should include the semantic representation of named landforms.

To achieve that aim in EcoLexicon regarding named bays, the knowledge should be represented in a semantic network according to the theoretical premises of Frame-based Terminology, which propose knowledge representations with explanatory adequacy for enhanced knowledge acquisition (Faber, 2009). Hence, each named bay should appear in the context of a specialized semantic frame that highlights both its relation to other terms and the relations between those terms. The construction of these semantic networks and the semi-automatic extraction of terms from a specialized corpus are described in this paper. As far as we know, this framework has not been studied in the context of specialized lexicography, which constitutes an original aspect of this work.

Distributional Semantic Models

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller & Charles, 1991). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance or cosine similarity (Salton & Lesk, 1968), *inter alia*.

Existing DSMs can be classified, based on two criteria, namely, the leveraged distributional information (Sahlgren, 2008), and the underlying language model (Baroni, Dinu & Kruszewski, 2014). According to the former criterion, models can be syntagmatic or paradigmatic.

Syntagmatic models capture combinatorial relations between terms, namely, non-hierarchical relations such as the effect of an entity on a process (e.g., ... *the Bay of Fundy, because of its basin geometry, amplifies tides*); where a process takes place (e.g., *Wind system changes affect also relative sea level as observed, for example, in the Hudson Bay*); or the location of an entity (e.g., *Many of the beaches along eastern Hudson Bay are characterized by boulder-strewn tidal flats*). Such syntagmatic relations are reflected in terms that co-occur within the same text region, either sentence, paragraph, or document

(Manning, Raghavan & Schütze, 2008). Latent Sematic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990) is an example of a syntagmatic model, whereby a term-document matrix of co-occurrences is first built to collect the normalized frequency of a term in a document, and the Singular Value Decomposition (Jolliffe, 2002) is then applied to reduce the number of columns to a few orthogonal latent dimensions.

Paradigmatic models are based on taxonomic relations such as hyponymy (e.g., *The Bay of Fundy is a low wave-energy environment that is dominated by tidal processes*) and meronymy (e.g., *Debris litters the bay floor along parts of the developed western shoreline of Greenwich Bay*). In these methods, a term-term matrix of co-occurrences indicates how many times context terms co-occur with a target term within a sliding context window, which spans a certain number of terms on either side of the target term. Hyperspace Analogue to Language (HAL) (Lund, Burges & Atchley, 1995) is an example of a paradigmatic model.

According to the second classification criterion, DSMs are either count-based or prediction-based. *Count-based models* calculate the frequency of the terms that occur within a term's context (i.e, a sentence, paragraph, document or context window of a certain size). LSA, HAL, Global Vectors (GloVe) (Pennington, Socher & Manning, 2014), and Correlated Occurrence Analogue to Lexical Semantic (COALS) (Rohde, Gonnerman & Plaut, 2006) are examples of this type of model. *Prediction-based models* exploit neural probabilistic language models, which represent terms by predicting the next term based on previous terms. Examples of predictive models include the continuous bag-of-words (CBOW) and skip-gram models (Mikolov, Chen, Corrado & Dean, 2013), Parallel Document Context (Sun, Guo, Lan, Xu, & Cheng, 2015), and Collobert and Weston model (Collobert & Weston, 2008).

The applications of DSMs in lexical and computational semantics include the following:

- Identification of semantic relations. DSMs are useful tools for Terminology, since they can help identify semantic relations between terms based on corpus data (Bertels & Speelman, 2014; Bernier-Colborne & L'Homme, 2015; Reimerink & León-Araúz, 2017). In addition, knowledge of a few seed terms and their relationships can help to infer analogous relationships for other similar terms that are nearby in the DSM (Hearst & Schütze, 1993; Widdows, 2003; Thompson, Batista-Navarro, Kontonatsios, Carter, Toon, McNaught, Timmermann, Worboys & Ananiadou, 2015).

- Information retrieval. Search engines can locate documents based on synonyms and related terms as well as matching keywords (Deerwester et al. 1990; Nguyen, Soto, Kontonatsios, Batista-Navarro & Ananiadou, 2017).

- Word sense discrimination and disambiguation. The vectors for each of the occurrences of the same term in a corpus (called context vectors) can be clustered, and the centroids of these clusters can be treated as word senses. An occurrence of the

same ambiguous term can then be mapped to one of these word senses, with a confidence level derived from the similarity between the context vector for this occurrence and the nearest centroids (Schütze, 1997 and 1998; Pantel & Lin, 2002).

- Use of word vectors as features for automatic recognition of named entities in text corpora (Turian, Ratinov, Bengio & Roth, 2009; El bazi & Laachfoubi, 2016), and for representation of proper names (Herbelot, 2015).

Clustering Analysis

Clustering is one of the most important unsupervised learning techniques in data analysis (Everitt, Landau & Leese, 2001). It classifies objects into groups (clusters) based on shared features. In hierarchical clustering, objects are successively integrated in inclusive clusters, depicted in dendrograms (Xu & Wunsch, 2009). Clustering techniques are used in many disciplines for purposes of Information Retrieval (Manning et al., 2008) and Text Mining (Feldman & Sanger, 2007), and, increasingly, in Corpus Linguistics (Moisl, 2009).

Work in lexical semantics that applies clustering techniques includes, *inter alia*, analysis of word distribution data in text to derive syntactic and semantic lexical categories (Bullinaria, 2008; Katrenko & Adriaans, 2008; Kiss, 1973; Miller, 1971); automatic induction of verb classes from verb selectional preferences extracted from corpus data (Sun & Korhonen, 2009); automatic metaphor identification in unrestricted text (Shutova, Sun & Korhonen, 2010); and classification of verbs into semantic groups based upon the relationship between words and grammatical constructions (Gries & Stefanowitsch, 2010).

Materials and Methods

Corpus Data

The terms related to named bays were extracted from a subcorpus of English texts on Coastal Engineering. This subcorpus, which comprises roughly 7 million tokens, is composed of specialized and semi-specialized texts, and is an integral part of the EcoLexicon English Corpus (23.1 million tokens) (see León-Araúz, San Martín and Reimerink [2018] for a detailed description).

GeoNames Geographical Database

The automatic detection of the named bays in the corpus was performed with a GeoNames database dump. GeoNames (http://www.geonames.org) has over 10 million proper names for 645 different geographical entities, such as bays, beaches, rivers, mountains, etc. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored. A daily GeoNames database dump is publicly available as a worldwide text file.

Pre-processing

After their compilation and cleaning, the corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased with the Stanford *CoreNLP* package for R programming language. The multiword terms stored in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

Named Bays Recognition

Both normalized and alternate names of the bays in GeoNames were searched in the lemmatized corpus. A total of 306 designations were recognized and listed. Nevertheless, since various designations can refer to the same bay because of syntactic variation (e.g., *Bay of Fundy* and *Fundy Bay*) and orthographic variation (e.g., *Choctaw[h]atchee Bay*), a procedure was created to identify variants and give them a single designation in the corpus.

In the case of syntactic variations without *of*, the preposition was automatically added to the names without it (e.g., *Fundy Bay* was converted to *Bay of Fundy*) and matched in the list of recognized designations. This was only problematic when the variants referred to different bays, such as the case of *Naples Bay* (USA) and *Bay of Naples* (Italy).

Orthographic variations were identified with a matrix of the Levenshtein edit distances between the 306 designations. The Levenshtein distance between two strings is the number of deletions, insertions, or substitutions required to transform the first string into the second one. As such, the pairs of strings with an edit distance of 1 or 2 were manually inspected to discover the orthographic changes.

Once the variants were normalized (Table 1) in the lemmatized corpus and joined with underscores, the number of named bays was 294. They are shown on the map in Figure 1, with color-coded rectangles that depict their frequency in the corpus. Their latitudes and longitudes were retrieved from the GeoNames database dump. This reflects the representativeness of the corpus in reference to bay locations and their number of mentions. As shown in Figure 1, most of the named bays are located in the USA.

Variant	Normalized designation
Paranague Bay	Paranagua Bay
Paranaguo Bay	Paranagua Bay
Choctawatchee Bay	Choctawhatchee Bay
Fundy Bay	Bay of Fundy
Funday Bay	Bay of Fundy
Ingleses Bay	Bay of Ingleses
Josiah's Bay	Josias Bay
Josiah Bay	Josias Bay
Westernport Bay	Western Port Bay
Port Phillip	Port Phillip Bay
Greenwich cove	Greenwich Bay
Halfmoon bay	Half Moon Bay

Table 1. Variants referring to the same bay and their normalized designation.
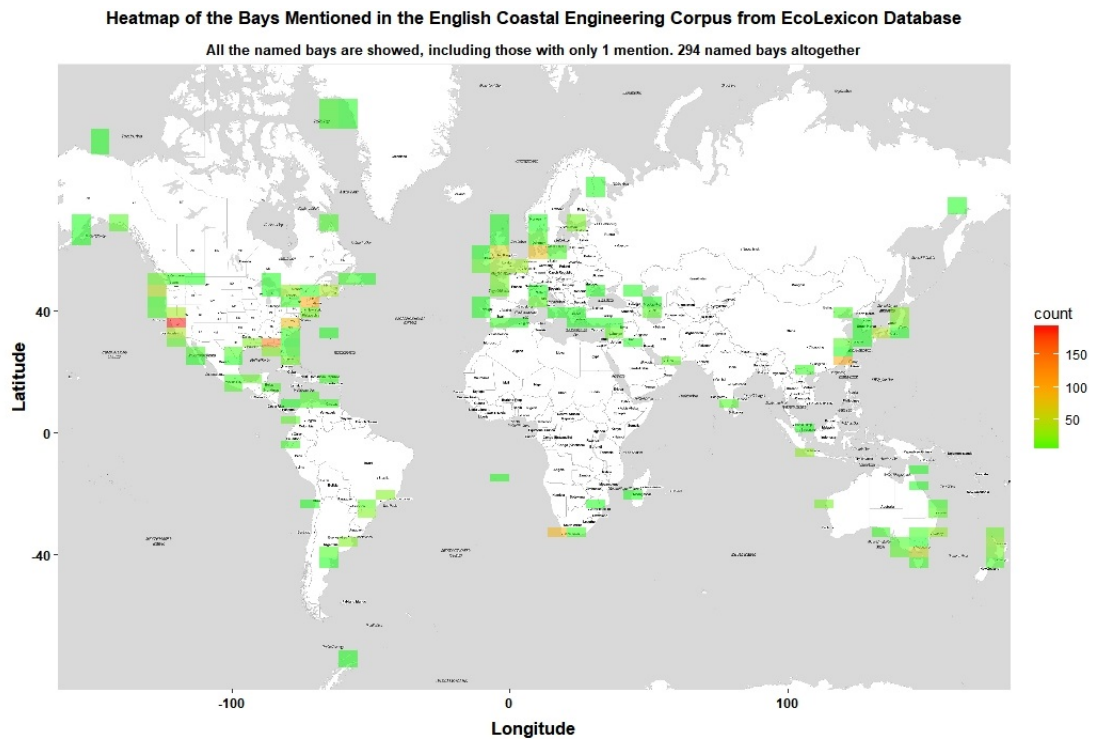
Figure 1. Map with the location and color-coded frequency of the 294 named bays.

A critical issue was the retrieval of the geographical coordinates of the bays. Although latitudes and longitudes could be retrieved from the GeoNames database dump, occasionally, the same designation referred to bays in different countries. For instance, the corpus only located *False Bay* in South Africa. However, GeoNames indicated that bays with the same name also existed in India, Yemen, the USA, Canada, and Australia. Such cases had to be resolved by corpus queries.

With regard to the occurrence frequency of the named bays in the corpus, the values ranged from 127 (*Monterey Bay*) to only one mention (150 of the 294 named bays). In our study, only those bays with an occurrence frequency greater than 5 were considered, since DSMs perform poorly with low-frequency terms (Luhn, 1957). Table 2 shows the 55 named bays that fulfilled this condition, whereas Figure 2 shows their number of mentions.

Country	Named Bays
United States (20)	**California State:** San Francisco Bay, Suisun Bay, Monterey Bay, San Diego Bay, Morro Bay, Back Bay. **Florida State:** Escambia Bay, Pensacola Bay, Tampa Bay, Saint Joseph Bay, Florida Bay. **State of New York:** Long Island Sound, Naples Bay. **State of Rhode Island:** Greenwich Bay, Narragansett Bay. **Other States:** Chesapeake Bay (Virginia), Siletz Bay (Oregon), Mobile Bay (Alabama), Delaware Bay (Delaware)
Australia (5)	**Victoria:** Port Phillip Bay, Western Port Bay, Apollo Bay.

	Other States: Botany Bay (New South Wales), Shark Bay (Western Australia)
United Kingdom (4)	**England:** Pevensey Bay, Start Bay, Morecambe Bay, Liverpool Bay
Japan (3)	Tosa Bay (Shikoku Island), Tokyo Bay (Kanagawa), Kamaishi Bay (Iwate)
Brazil (2)	Sepetiba Bay (Rio de Janeiro), Imbituba Bay (Santa Catarina)
Canada (2)	Bay of Fundy (Nova Scotia), Hudson Bay (Ontario)
France (2)	Baie des Anges (Provence-Alpes-Côte d'Azur), Baie des Veys (Normandy)
Mexico (2)	Bay of Campeche (Campeche), Todos Santos Bay (Baja California)
New Zealand (2)	**Auckland:** Bay of Plenty, Tauranga Harbor
South Africa (2)	**Western Cape:** False Bay, Gordons Bay
The Netherlands (2)	**Wadden Sea:** Ley Bay, Dollard Bay
Argentina (1)	Samborombon Bay (Buenos Aires)
China (1)	Quanzhou Bay (Fujian)
Colombia (1)	Buenaventura Bay (Valle del Cauca)
Denmark (1)	Kogo Bay (Zealand)
Estonia (1)	Tallinn Bay (Harjumaa)
Indonesia (1)	Jakarta Bay (Jakarta)
Iran (1)	Chabahar Bay (Sistan and Baluchestan)
Ireland (1)	Dingle Bay (Munster)
Spain (1)	Bay of Biscay (Basque Country)

Table 2. Designations and locations of the 55 named bays whose occurrence frequency was higher than 5.

Figure 2. Designations and number of mentions of the 55 named bays whose occurrence frequency was higher than 5.

Term-term matrix construction

After the 294 named bays were joined with underscores in the lemmatized corpus, a count-based DSM was built with the R package *quanteda* for text mining. A count-based DSM w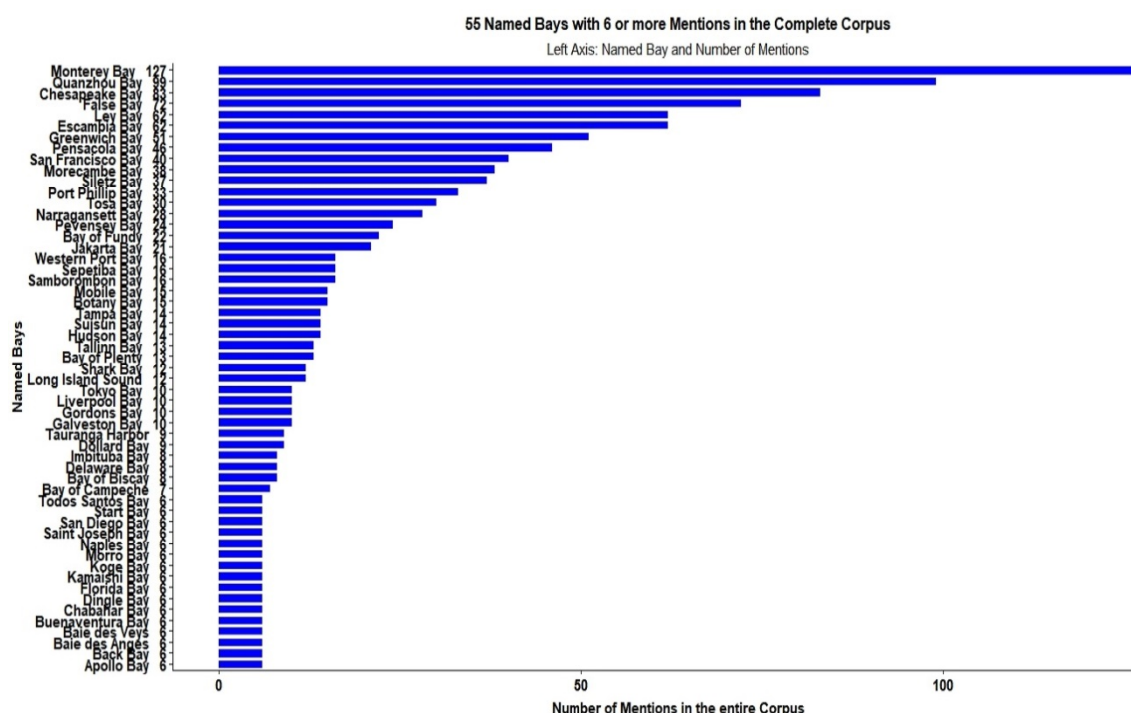as selected to obtain term vectors since this type of DSM outperforms prediction-based ones on small-sized corpora of under 10 million tokens (Ars, Willits & Jones, 2016; Sahlgren & Lenci, 2016).

In the DSM, only terms larger than 2 characters were considered, and numbers and punctuation marks were removed. Additionally, the minimal occurrence frequency was set to 5 so that the co-occurrences were statistically reliable (Evert, 2007). A sliding context window was set up to span 20 terms on either side of the target term because for small corpora, large windows lead to larger counts and greater statistical reliability (Rohde et al., 2006, p. 31; Bullinaria & Levy, 2007, p. 522). Furthermore, when the window is larger, the relations in the DSM will be more semantic than syntactic (Jurafsky & Martin, 2017, p. 5). Since closed-class words are often considered too uninformative to be suitable context words (Kiela & Clark, 2014), stopwords, adjectives and adverbs were not used as context words.

The resulting DSM was a 4,431×4,431 frequency matrix *A*, whose row vectors represented the 55 named bays plus the 4,376 different terms inside the context windows of 20 terms on either side of those bays.

Selection of bays and terms for clustering purposes

Subsequently, a 55×4,376 submatrix *B* was extracted from *A*, where the rows represented the 55 named bays, and the columns represented the 4,376 terms co-occurring with the bays. To cluster the bays of *B* sharing the same associated terms, it was necessary to select both the bays and the terms that best discriminated different groups of bays. This was done by removing the bays and the terms that could act as random noise and adversely affect the clustering results (Kaufman & Rousseeuw, 1990). The remainder of this section explains the selection method of bays and terms for clustering purposes.

An issue often highlighted in the literature on the clustering of rows in a frequency matrix abstracted from corpus data is that variation in document length will affect the clustering results. These documents are thus clustered in accordance with relative length rather than with a more interesting latent structure in the data (Moisl, Maguire & Allen, 2006; Rojas-Garcia, Faber & Batista-Navarro, 2018; Thabet, 2005). The conventional solution to the problem is to normalize the values in the frequency matrix to mitigate the effect of length variation. Normalization by mean document length (Spärck, Walker & Robertson, 2000) is widely used in Information Retrieval literature.

Nevertheless, as stated by Moils (2011), there is a limit to the effectiveness of normalization, and it has to do with the probabilities with which the terms in the column vectors occur in the corpus. Some documents in the matrix rows might be too short to give

accurate population probability estimates for the terms, and since length normalization methods accentuate such inaccuracies, the result is that analysis based on the normalized data inaccurately clusters the rows. One solution consists in statistically ascertaining which documents are too short to provide good estimates and to remove the corresponding rows from the matrix.

For that aim, Moisl (2011, pp. 42-45) proposes a formula that calculates the document length necessary to estimate the probability of each term in the column vectors with a 95% confidence level. Therefore, the formula can be applied to establish a minimum length threshold for the documents and to eliminate any documents under that threshold.

In our case, a document was considered to be the set of all context windows where a certain named bay appeared, and thus corresponded to a row of matrix *B*. As such, we had 55 named-bay documents. Similarly, the length of a document was considered to be the total number of words appearing in the set of all context windows of a certain named bay. The document lengths ranged from 4,950 words (for *Monterey Bay*) to 232 words (for *Kamaishi Bay*). Moisl's (2011) method was then applied to matrix *B* to determine: (1) which of the 55 named bays should be eliminated from our analysis; and (2) which terms helped to distinguish different groups of the retained bays.

Table 3 shows the length for named-bay documents needed by each of the 4,376 terms in the columns of matrix *B* so that their population probabilities could be estimated with a 95% confidence level, according to Moisl's (2011) formula. The terms in Table 3 were sorted in ascending order of the required document length.

Index	Term	Length needed for named-bay documents
1	**beach**	**416**
2	sea_surface_temperature	475
3	island	530
4	river	574
5	bay	597
6	wave	644
7	hurricane	655
[…]	[…]	[…]
325	la_niña	4,927
326	natural_area	4,942
327	criterion	4,944
328	**canal**	**4,944**
329	spring	4,952
330	pass	4,957
331	season	4,968
332	organic_material	4,975
[…]	[…]	[…]
4,371	swash_flow	522,884

4,372	morphologic_change	522,884
4,373	locally_generated_wave	522,884
4,374	counter-circulation	522,884
4,375	rip-opposite_megacusp	522,884
4,376	rip_current_experiment	522,884

Table 3. Length needed for named-bay documents (mostright column) associated with each of the 4,376 terms (middle column) co-occurring with the bays, according to Moisl's (2011) formula.

Since the lowest document-length value needed by the terms was 416 words (for the term *beach* in the first row of Table 3), those bays whose document length was smaller than the minimum length threshold 416 were eliminated from the analysis. This meant that only 29 bays of 55 were retained. As expected, the 29 named bays selected by Moisl's (2011) method were those with the highest number of mentions in the corpus (Figure 2), from Monterey Bay (127 mentions) to Shark Bay (12 mentions).

Regarding the selection of terms, since the maximum length of our named-bay documents was 4,950 words, only the first 328 terms in Table 3 were retained for clustering purposes because their needed document lengths were less than 4,950 words. These results are plotted in Figure 3, where the 4,376 terms co-occurring with the 55 bays are on the horizontal axis (sorted in ascending order of the needed document length), and their required document lengths are on the vertical axis. The red horizontal line indicates the maximum length of the named-bay documents (4,950 words), and the green vertical line marks the 328 terms whose needed document lengths were equal to or less than the maximum named-bay document length.
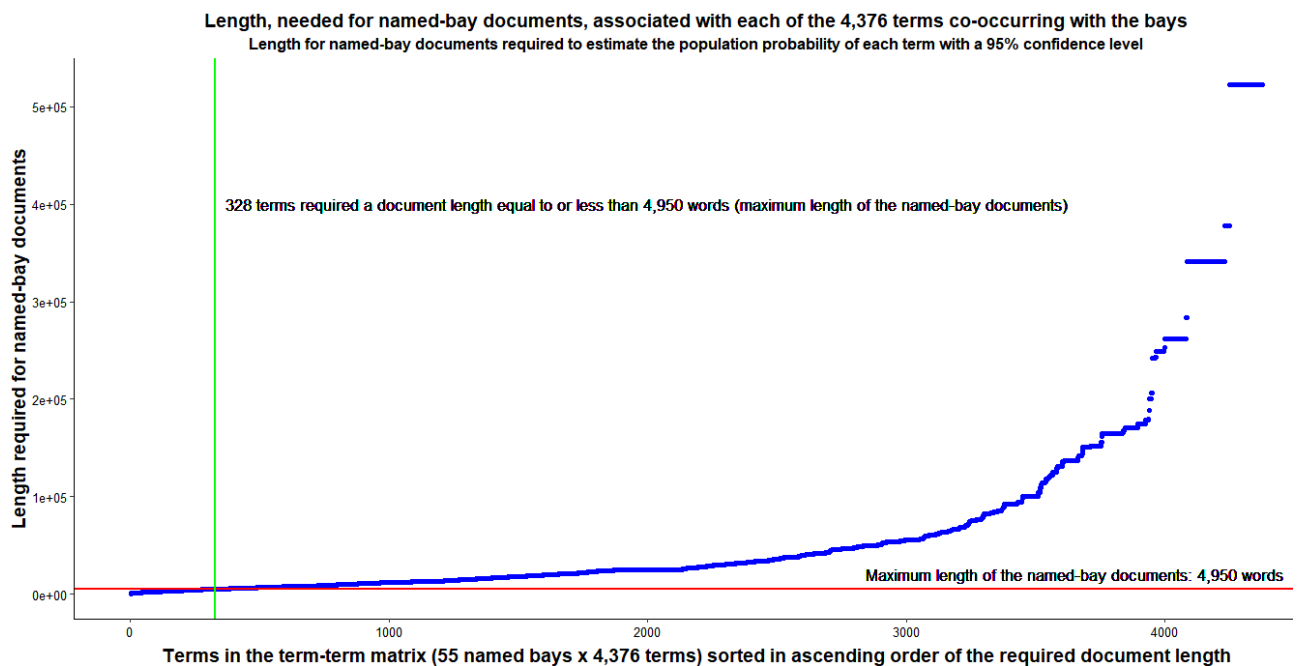
Figure 3. The required document lengths (vertical axis) associated with each of the 4,376 terms (horizontal axis) co-occurring with the 55 named bays.

Of the 328 terms selected by Moisl's (2011) method, only 310 terms co-occurred with the 29 retained bays. Therefore, a 29×310 submatrix *C* was extracted from *B* to group the bay vectors. A visualization of the 310-dimensional bay vectors in a 2-dimensional space is shown in Figure 4. This was accomplished by first weighting the frequency matrix *C* by using the *log-likelihood* association measure (see following section), reducing the number of dimensions via Singular Value Decomposition (Jolliffe, 2002), and plotting the data points according to the first two principal-component coordinates.
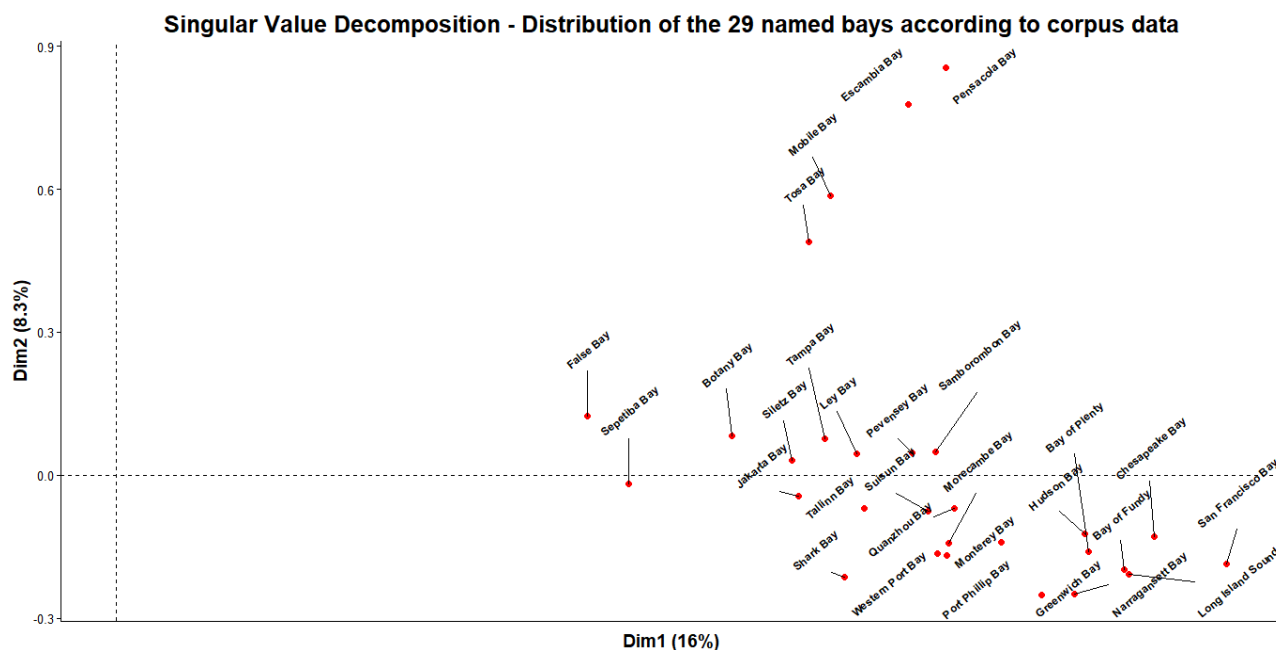
Figure 4. Visualization of the 29 bay vectors in a 2-dimensional space.

Clustering of named bays and weighting schemes

According to Moisl (2011, pp. 30-31), the 29×310 frequency matrix *C* was first normalized by mean document length. Next, we applied a hierarchical clustering technique, using the squared Euclidean distance as the intervector distance measure and Ward's Method as the clustering algorithm (Xu & Wunsch, 2009).

Since it is not clear how strong a cluster is supported by data (Suzuki & Shimodaira, 2004), a means for assessing the certainty of the existence of a cluster in corpus data was devised. Multiscale bootstrap resampling (Shimodaira, 2004) is a method for this in hierarchical clustering, which is implemented in the R package *pvclust* (Suzuki & Shimodaira, 2006). For each cluster, this method produces a number ranging from zero to one. This number is the approximately unbiased probability value (AU *p*-value), which represents the possibility that the cluster is a true cluster. The greater the AU *p*-value, the greater the probability that the cluster is a true cluster supported by corpus data. An AU

p-value equal to or greater than 95% significance level is most commonly adopted in research.

In the clustering results, 2 groups of bays, with AU *p*-values equal to or greater than 95%, were considered (Figure 5). Unfortunately, the existence of only 2 groups with such a large number of bays inside each of the clusters was not conducive to appropriately describing the semantic frames of the bays. As such, the normalization by mean document length was disregarded because it led to unreliable clustering. Consequently, other weighting schemes were tested instead.
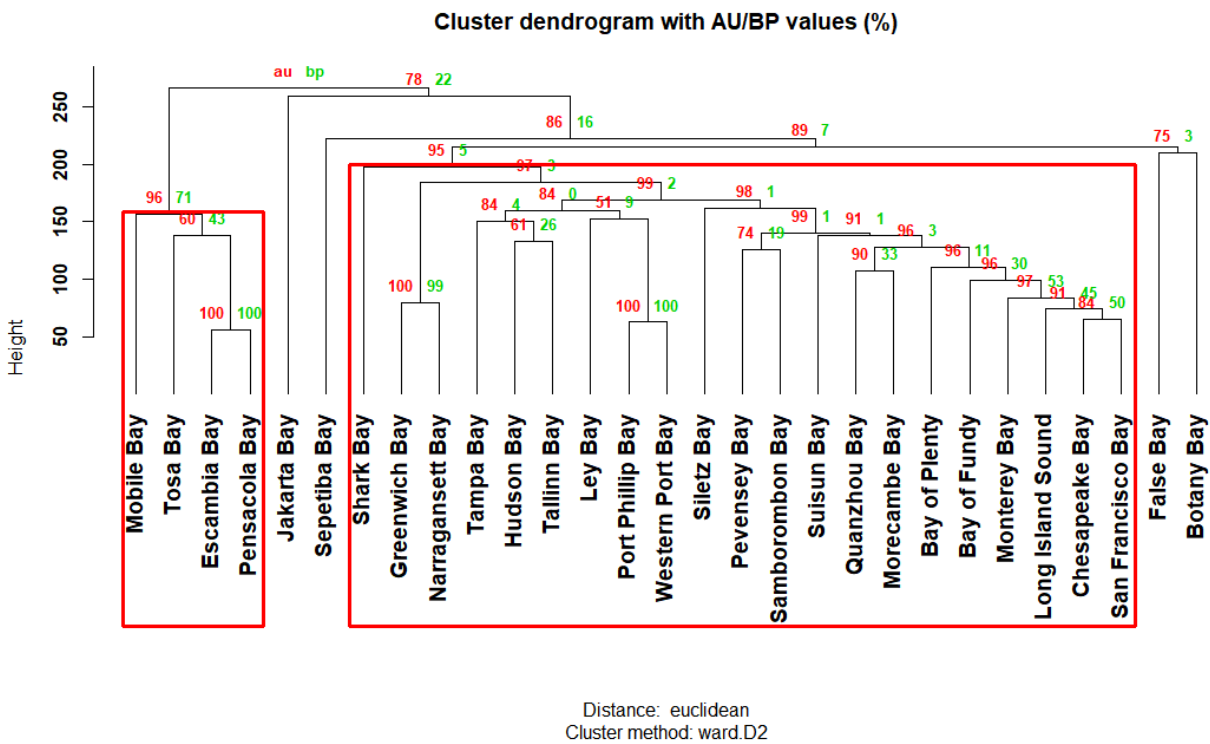
Figure 5. Dendrogram of the hierarchical clustering of the 29 named bays, along with 2 red rectangles indicating clusters with red-colored AU *p*-values ≥ 95% (red values at branches). The matrix was previously normalized by mean document length, which led to unreliable clustering results.

The frequency matrix *C* was subjected to three weighting schemes. First, the statistical *log-likelihood* measure (Dunning, 1993) was applied to calculate the association score between all term pairs, including the named bays (Evert, 2007, pp. 24-30). Research on computational linguistics reveals that *log-likelihood* is able to capture syntagmatic and paradigmatic relations (Bernier-Colborne & Drouin, 2016, p. 58; Lapesa et al., 2014, p. 168) and to achieve better performance for medium-to-low-frequency data than other association measures (Alrabia et al., 2014, p. 4; Krenn, 2000). However, the calculation of the *log-likelihood* scores was modified to cope with these critical situations:

- When the observed frequency was less than the expected one, the score was set to 0, as recommended by Evert (2007, p. 22). Otherwise, the score would have been

negative showing repulsion between terms, whereas our interest was in the stronger attraction to each other.

- When a term pair did not co-occur (i.e., its observed frequency was 0), the score was set to 0. Otherwise, the score would have obtained a low value, indicating a certain attraction between the pair of terms despite the absence of co-occurrence in corpus data.
- When a term co-occurred with only one bay, the corresponding addend in the *log-likelihood* formula (i.e., the addend where the observed frequency O_{21} takes part, according to Evert [2007, p. 25]) was set to 0. Otherwise, the score would have tended to minus infinity, and its value would have been undetermined.

Secondly, the association scores were transformed by adding 1 and calculating the natural logarithmic to reduce skewness (Lapesa et al., 2014). Finally, the row vectors were normalized to unit length to minimize the negative effects of extreme values on the Euclidean distance-based clustering technique.

The hierarchical clustering technique was then applied to the weighted matrix *C*. As a result, 5 groups of bays with AU *p*-values equal to or greater than 99% were considered to be strongly supported by corpus data (Figure 6). Two bays comprised each of the 5 clusters, which provided evidence that the clustering results with a *log-likelihood* measure was more reliable than those with mean document length. Accordingly, this paper focuses on the 10 named bays inside the 5 clusters shown in Figure 6.

Cluster dendrogram with AU/BP values (%)

Distance: euclidean
Cluster method: ward.D2
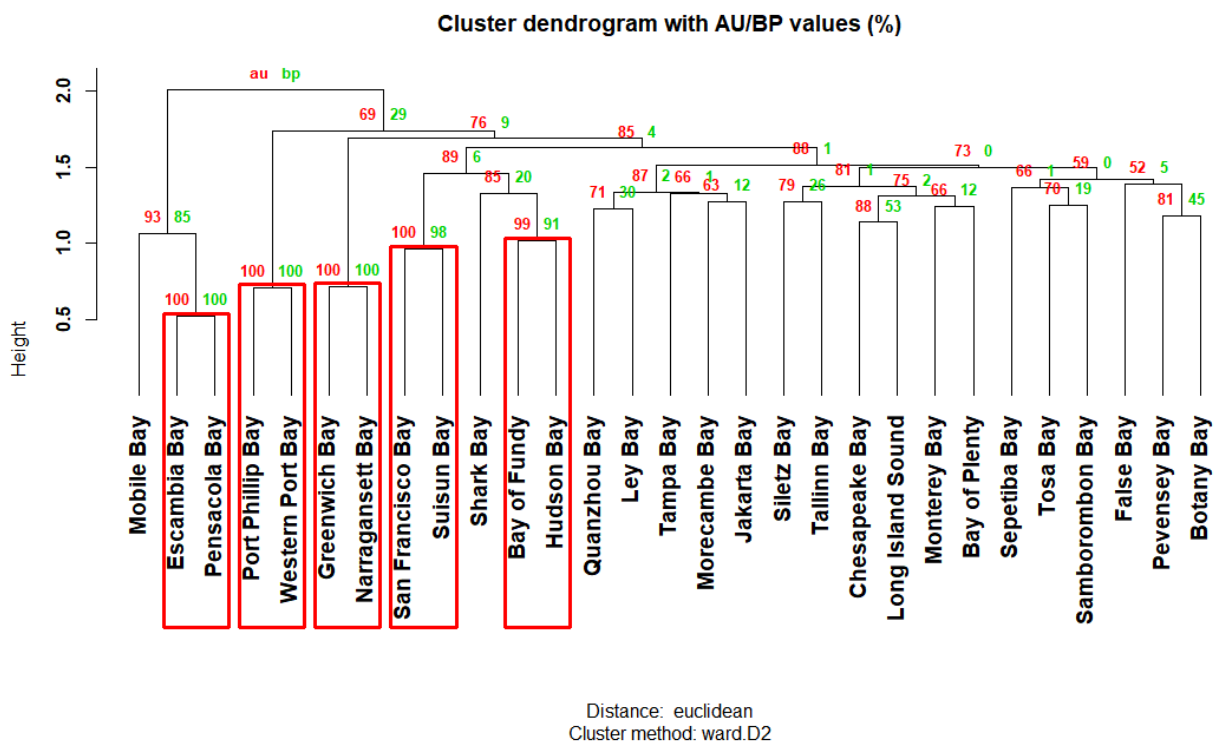
Figure 6. Dendrogram of the hierarchical clustering of the 29 named bays, along with 5 red rectangles indicating clusters that are strongly supported by corpus data (red-colored AU *p*-values ≥ 99%). The matrix was previously weighted by *log-likelihood* measure, which led to reliable clustering results.

Figure 7 shows a scatter plot of these 5 clusters via Singular Value Decomposition. The weighted matrix *C* was also used to visualize the bay vectors in Figure 4.
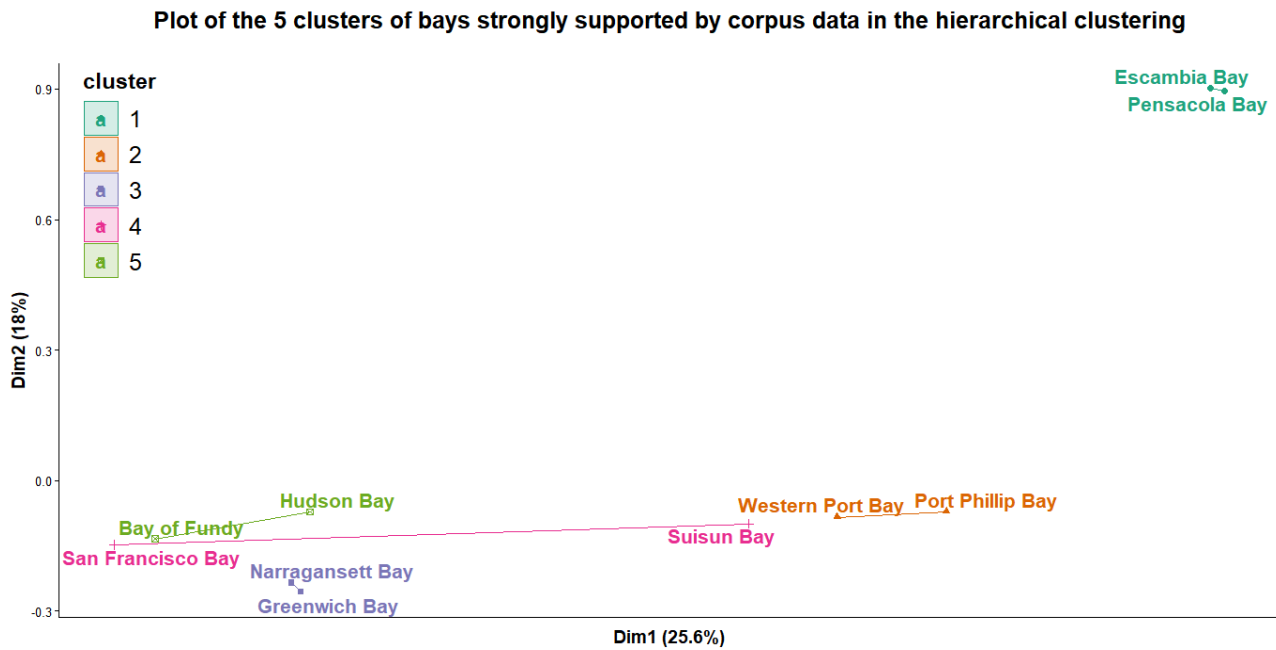
Figure 7. Scatter plot of the 5 clusters of bays, strongly supported by corpus data, via Singular Value Decomposition.

Selection of terms for semantic network construction

With a view to evaluating the procedure for term selection that best captured the terms related to the 29 named bays for the construction of semantic networks, 5 methods were devised.

Method One.

A 339×339 squared frequency matrix *D1* was built, whose rows represented the 29 named bays plus the 310 terms selected by Moisl's (2011) method. The columns also represented the same bays and terms co-occurring with the target words in the rows. *D1* was weighted by the *log-likelihood* measure. Then, the scores were transformed by adding 1 and calculating the natural logarithmic to reduce skewness (Lapesa et al., 2014).

The matrix *D1* tested whether the 310 terms selected by Moisl's (2011) method were sufficient to understand and represent the semantic frames in which the 29 named bays appeared.

Method Two.

A 3.867×3.867 frequency matrix *D2* was built, whose rows represented the 29 named bays plus the 3.838 terms co-occurring with them. *D2* was weighted in the same way as *D1*. *D2* tested whether no term selection method could optimally describe the semantic frames of the bays.

Method Three.

The number of columns in the weighted matrix *D2* was reduced to only two by applying the innovative dimensionality reduction technique UMAP (Uniform Manifold Approximation and Projection) (McInnes & Healy, 2018). It eliminates information redundancy among column variables and helps to identify local latent structures in corpus data. As a result, a 3.867×2 matrix *D3* was obtained.

D3 was tested whether such an innovative dimensionality reduction technique applied to all the terms co-occurring with the 29 bays was an improvement over *D2*.

Method Four.

In the same way as Moisl's (2011) method was used to select bays and terms, another statistical method was employed to select the terms that best described the 29 bays, based on Moisl (2015, pp. 77-93). In Corpus Linguistics, Moisl (2015) suggests retaining the term columns with the highest values in four statistical criteria: *raw frequency*, *variance*, *variance-to-mean ratio* (vmr) and *term frequency-inverse document frequency* (tf-idf).

Moisl's (2015) method was applied to a 29×3.838 frequency matrix, whose rows represented the 29 named bays. The columns represented all the terms co-occurring with them (excluding the bays). Figure 8 shows the co-plot of the four criteria, *z*-standardized for comparability reasons, and sorted in descending order of magnitude. A threshold of up to 1000 was set. This meant that only 847 terms fulfilled all criteria.

We estimated that between 25 and 30 terms would be necessary for a named bay to describe its semantic frame. A total number of terms ranging from 725 to 870 would thus be required for the description of the 29 bays. The threshold was set accordingly, so that the number of selected terms was within the interval 725-870 terms.
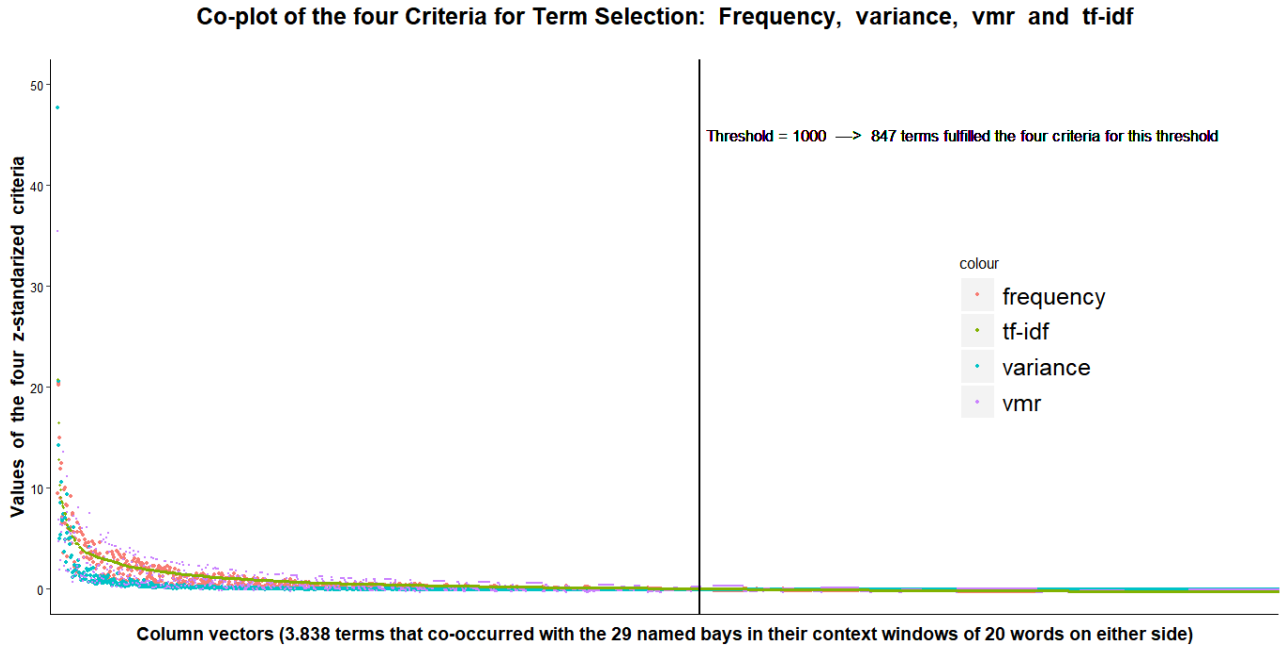
Figure 8. Co-plot of the 4 criteria for term selection: Frequency, variance, vmr, and tf-idf.

An 876×876 frequency matrix *D4* was obtained, where the rows represented the 29 named bays plus the 847 terms selected by Moisl's (2015) four statistical criteria. *D4* was weighted in the same way as *D1*.

D4 tested whether a term selection method was needed to appropriately describe the semantic frame of a named bay.

Method Five.

Finally, the 876 columns in the weighted matrix *D4* were reduced to only two columns by applying the UMAP technique. As a result, an 876×2 matrix *D5* was obtained. *D5* tested whether dimensionality reduction by UMAP, applied to selected terms, was an improvement over *D4*.

Terms characterizing each cluster

To ascertain the terms closely associated with each of the 5 clusters for semantic network construction, the following procedure was used:

1. For each of the 10 named bays in the clusters (Figure 6), a set of the top-30 terms, most semantically related to each bay according to their cosine similarities, was extracted from the corresponding DSM.
2. For each cluster, the mathematical operation *set intersection* was applied to the sets of the top-30 terms most semantically related to both bays in the same cluster. Only the shared terms with a cosine similarity higher than 0.4 were selected.

A reduced set of terms was thus obtained for each cluster to describe the named bays, based on shared associated terms.

Results

Analysis of the term selection methods

For each of the clusters, the term selection methods produced 5 sets of terms, which characterized them. Those term sets were qualitatively compared to gold standard sets of terms, manually extracted from the context windows of the 10 bays clustered in Figure 6, which best described each of the clusters for semantic network construction. For space constraints, only the main results of the comparisons are highlighted:

- The method that systematically produced the best sets of terms characterizing each cluster for semantic network construction was the Method Four, consisting in term selection based on Moisl's (2015) four statistical criteria.
- Method One, which consisted of bay and term selection according to Moisl (2011), produced sets of terms that could be used to infer the scientific topic in which the bays of a cluster were involved. However, because of the small number of terms selected (310 terms for 29 bays), the term sets were not conducive to understandable knowledge representations because, most of the time, the number of terms was not sufficient to derive a clear semantic relation between them.
- Method Two, with all the 3,838 terms co-occurring with the bays, produced meaningful sets of terms, which were suitable for semantic network construction. Nonetheless, surprisingly, the number of terms in the sets was lower than that in the sets obtained with Method Four. The reason was that the number of shared terms in each cluster with a cosine similarity higher than 0.4 was lower with Method Two, and higher with Method Four. In addition, for most clusters, some terms in the sets obtained with Method Two were not relevant for frame construction.
- Method Three and Method Five, whereby the number of columns was reduced to two with the UMAP technique, produced unreliable term sets. Firstly, both methods selected some terms with low occurrence frequency that could be disregarded for frame description. Secondly, they selected certain terms that were related to only one of the bays in a cluster. Thirdly, both methods selected some terms that were not related to any of the bays in a cluster. Those unrelated terms were associated with some of the terms that were directed related to the bays in a cluster, but in a thematic context different from that in which the bays were involved.

For the construction of the semantic frames presented in the next section, Method Four was thus applied.

Semantic frames describing the bay clusters
Interestingly, the five clusters in Figures 6 and 7 contained bay pairs located in the same geographical areas, as shown in Table 4.

Cluster 1 (USA)	Cluster 2 (Australia)	Cluster 3 (USA)	Cluster 4 (USA)	Cluster 5 (Canada)
Escambia Bay (Florida)	**Port Phillip Bay** (Victoria)	**Greenwich Bay** (Rhode Island)	**San Francisco Bay** (California)	**Bay of Fundy** (Nova Scotia)
Pensacola Bay (Florida)	**Western Port Bay** (Victoria)	**Narragansett Bay** (Rhode Island)	**Suisun Bay** (California)	**Hudson Bay** (Ontario)

Table 4. Designations and locations of the bays in the 5 clusters.

For the description of the frames, the semantic relations were manually extracted by querying the corpus in Sketch Engine (Kilgarriff, Rychly, Smrz & Tugwell, 2004), and analysing knowledge-rich contexts (Meyer, 2001). The query results were concordances of any elements between the bays in a cluster and related terms in a ±30 span. The semantic relations were those in EcoLexicon (Faber, León-Araúz & Prieto, 2009), with the addition of *does_not_affect*, *not_located_at*, *increases*, *decreases*, *belongs_to*, *uses*, *simulates*, and *becomes*.

In the first cluster, *Escambia* and *Pensacola* bays are thematically related by numerical parameter studies that simulate: (1) hurricane-induced storm surges, waves and winds, and the land dissipation effect on wind; (2) the effects of these features and inlet-bay configuration on open-coast storm-surge hydrographs. To validate simulation results, researchers employ historical data of the effects of *Hurricane Ivan* on both bays. Figure 9 shows the terms highly associated with the bays and their semantic relations.

Figure 9. Semantic network of the terms associated with the *Escambia* and *Pensacola* bays.

The bays in the second cluster are involved in the topic of Integrated Coastal Management (ICM). Since the environmental condition of the Victorian coast (Australia) has not improved despite thirty years of ICM, case studies have been carried out in different coastal environments located on the *Port Phillip* and *Western Port* bays: a coastal headland (Point Nepean), a coastal lakes system (Gippsland Lakes), and an urbanising coastal region (Geelong region). These environments were examined to develop an approach that incorporates ICM in a Sustainable Coastal Planning, which responds to the pressures of urban growth, tourism, decline in water quality, climate change on coasts, coastal planning, and environmental protection (Figure 10).
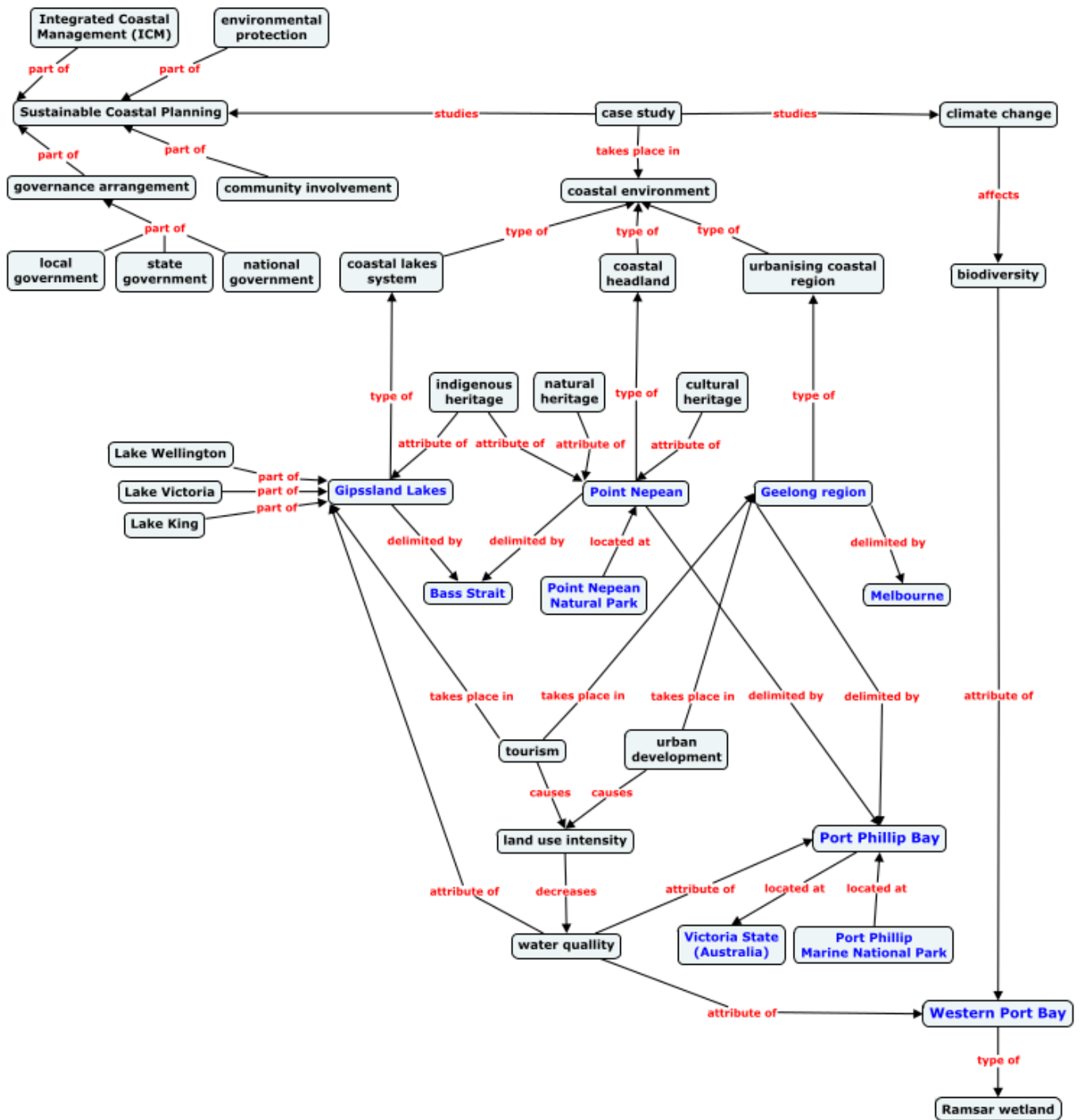
Figure 10. Semantic network of the terms associated with the *Port Phillip* and *Western Port* bays.

In the third cluster, *Greenwich* and *Narragansett* bays are sites for the study of benthic geologic habitats, namely, spatially recognizable areas in bay floors with special geologic and biologic characteristics. These habitats are identified by using imagery, and then classified according to criteria such as sediment particle size (Figure 11).
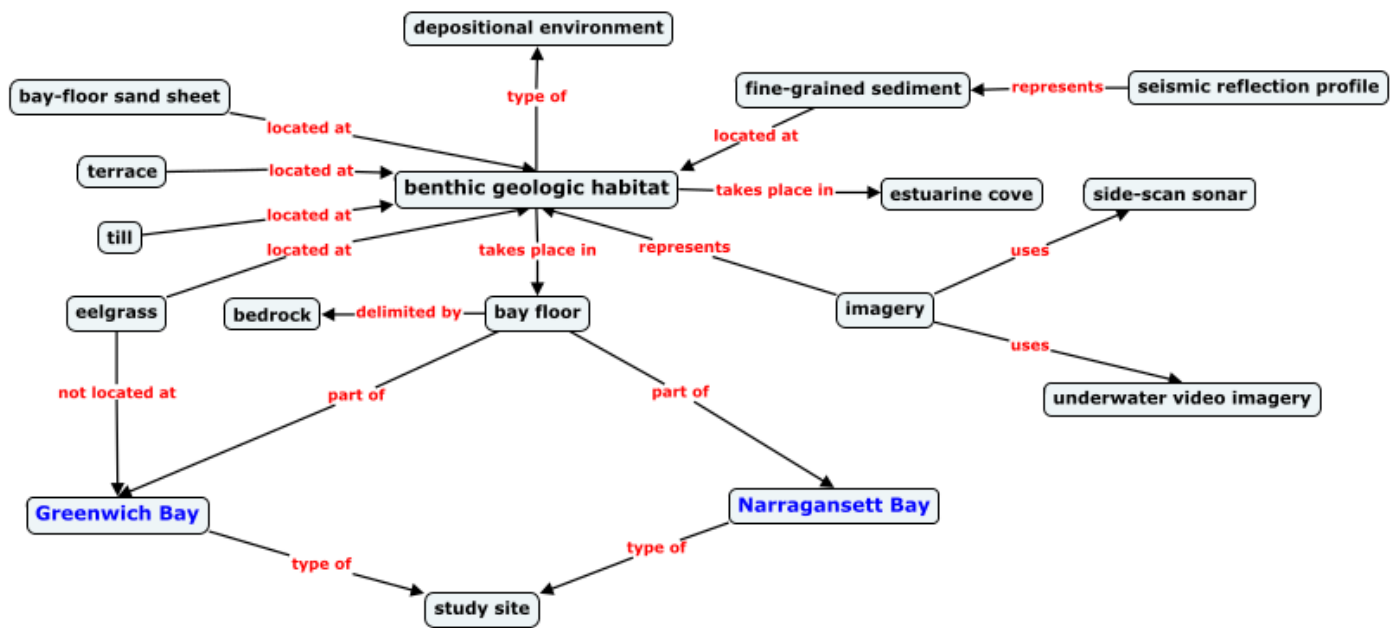
Figure 11. Semantic network of the terms associated with the *Greenwich* and *Narragansett* bays.

In the fourth cluster, *San Francisco* and *Suisun* bays are involved in research studies to determine whether the timescale dependence of forcing mechanisms on suspended sediment concentration (SSC) is typical in estuaries, based on SSC data. Of the forcing mechanisms, several tend to be concurrently active in estuaries, rather than only one. Multiple active forcing mechanisms have been observed in estuaries, but the specific mix of active mechanisms is different in each (Figure 12).
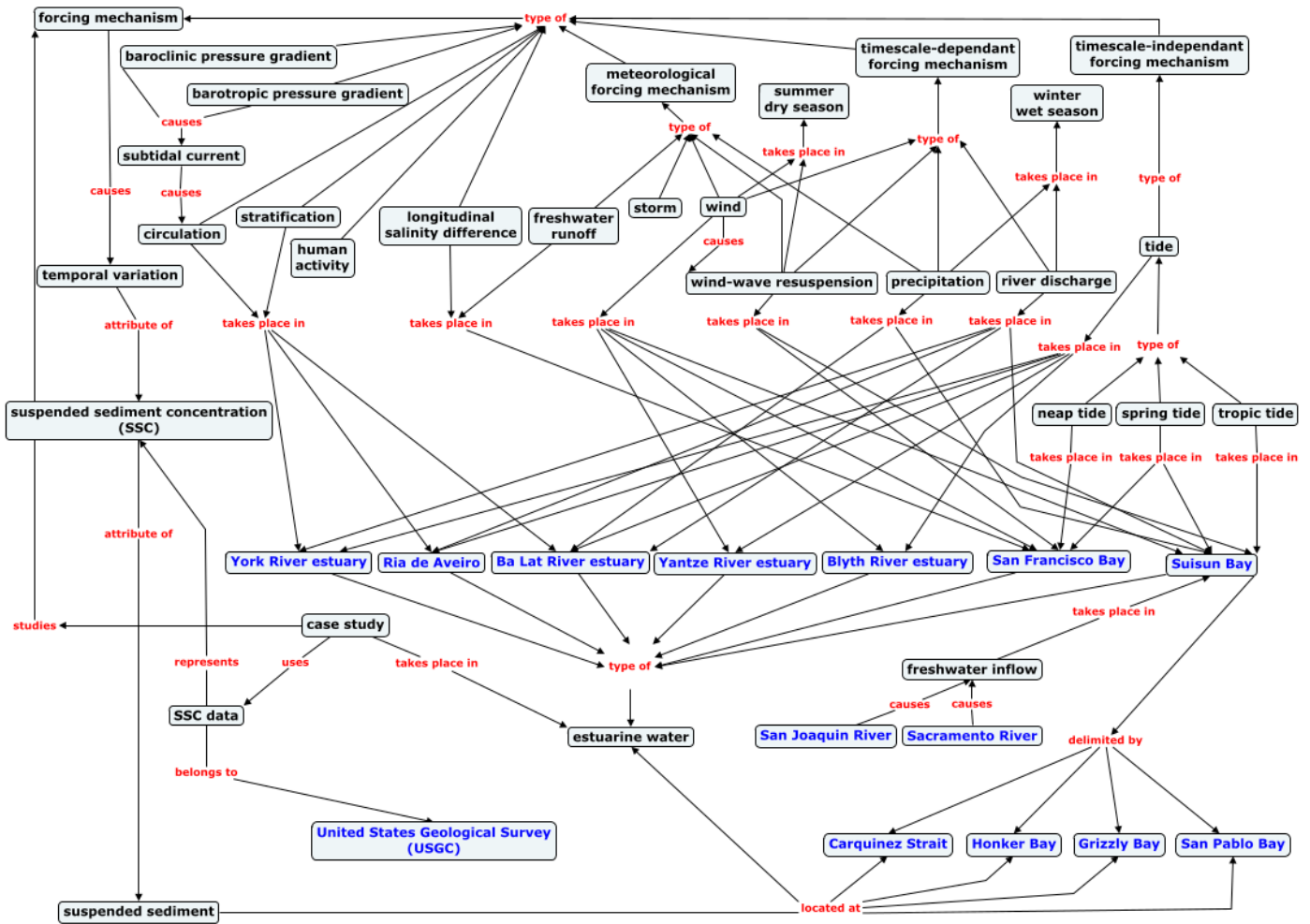
Figure 12. Semantic network of the terms associated with the *San Francisco* and *Suisun* bays.

Finally, in the fifth cluster, *Bay of Fundy* and *Hudson Bay* are low wave-energy environments with large sedimentation rates and tidal ranges, which originate tidal flats and tidal marshes. The *Bay of Fundy* is a vertically mixed estuary. With limited freshwater inputs and the largest tidal ranges in the world (over 15 meters), it is used to generate electricity, thanks to a Straflo turbine. These strong tides also erode joint planes (vertical cracks) of cliffs on the bay. As a result, joint planes enlarge and become caves, which erode further and form arches. When the roof of these arches collapses, the stacks on the bay are formed (Figure 13).
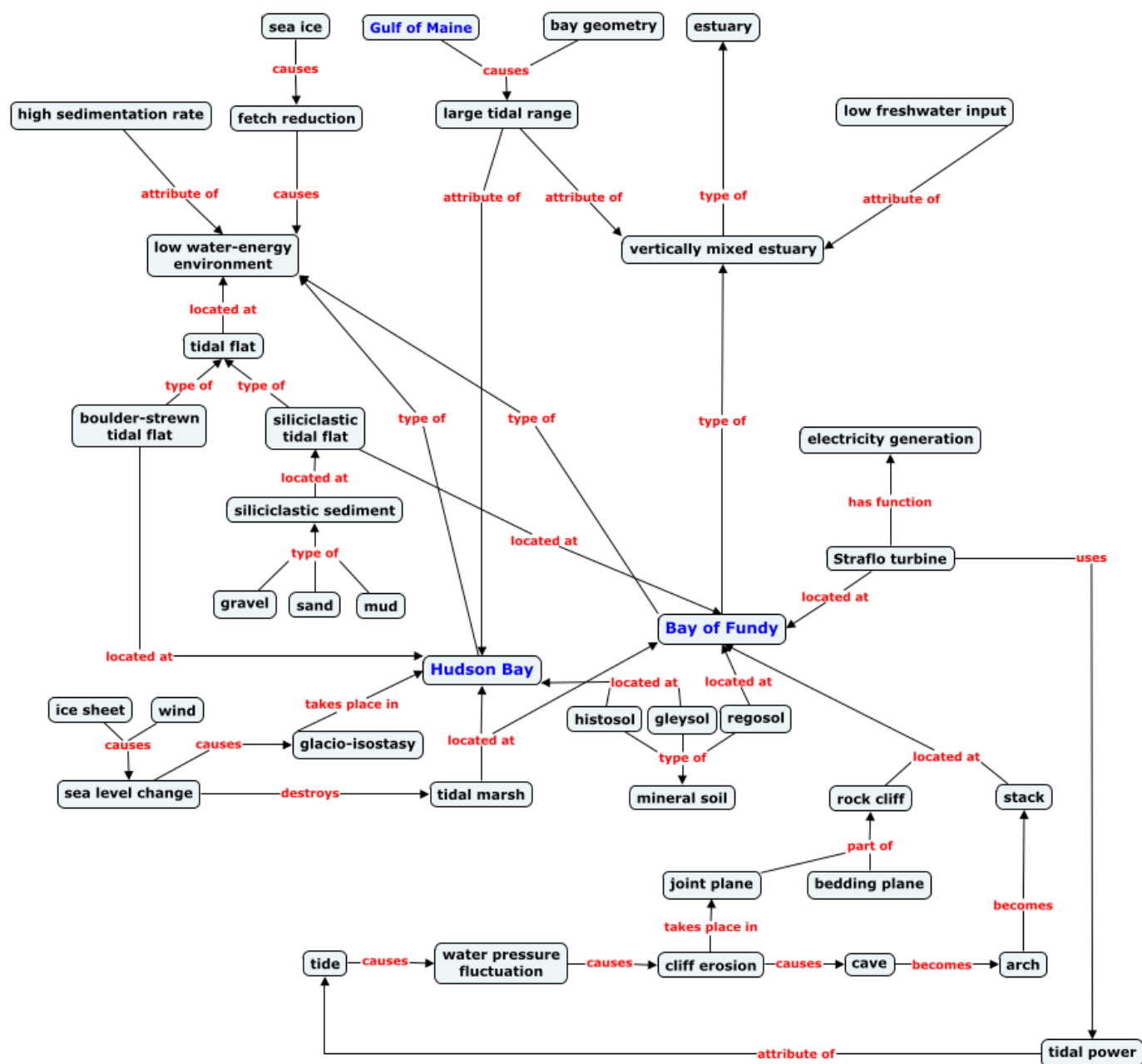
Figure 13. Semantic network of the terms associated with *Bay of Fundy* and *Hudson Bay*.

Conclusions

To extract knowledge for the semantic frames or conceptual structures (Faber, 2012) that underlie the usage of named bays in Coastal Engineering texts, a semi-automated method for the extraction of terms and semantic relations was devised. The semantic relations linking concepts in the semantic frames were manually extracted, based on the corpus analysis of knowledge-rich contexts (Meyer, 2001), a time-consuming task that is essential for the explanatory adequacy of frames (Faber, 2009). In future research, the knowledge patterns by León-Araúz, San Martín & Faber (2016) for the automatic extraction of semantic relations will be tested.

The method for the extraction of terms closely associated with named bays combined selection procedures for both terms and bays, with the use of a count-based DSM, weighted by a *log-likelihood* association measure. The selection of 29 named bays from an initial set of 55 bays with an occurrence frequency greater than 5 was performed by using Moisl's (2011) statistical method. It consisted in determining which bays had suitable document lengths for accurate estimation purposes. This bay selection procedure, along with a matrix normalization by *log-likelihood* measure, yielded reliable clustering results when the bays were automatically grouped based on their shared terms. Surprisingly, the normalization by mean document length, widely used in Information Retrieval, and suggested by Moisl (2011) because of its intuitive simplicity, did not achieve the desired clustering results. This reinforces the view that the performance of conventional procedures used in Natural Language Processing (NLP) largely depends on the nature of the task.

Regarding the term selection procedures, of the five methods tested, that of Moisl (2015, pp. 77-93), based on four statistical criteria, obtained the best performance for semantic network construction when qualitatively compared with gold standard sets of terms. Nonetheless, for reliable bay clustering, the best term selection procedure was that of Moisl (2011). This finding reveals that the best set of terms characterizing named bays is different, depending on whether the ultimate goal is clustering or frame description.

The two methods for term selection including dimensionality reduction by UMAP produced poor results. Since the reduction to two dimensions was probably insufficient, a larger number of dimensions will be tested in the future. Moreover, Topic Modelling (Blei et al., 2003), a domain-specific dimension reduction technique for texts, will be also applied.

Finally, the semantic frames in the previous section reflect that most terms related to named bays are multiword terms (MWT) since specialized language units are mostly represented by such compound forms (Nakov, 2013). The MWT extraction was possible because they were previously matched and joined with underscored in the lemmatized corpus, thanks to the list of MWTs stored in EcoLexicon. This implies that EcoLexicon is a valuable resource for any NLP tasks related to specialized corpora on environmental science.

Acknowledgements

Footnotes
2. http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi
3. https://www.eionet.europa.eu/gemet/en/themes/
4. http://www.fao.org/faoterm/en/

5. http://aims.fao.org/en/agrovoc
6. http://www.environmentontology.org/Browse-EnvO

References

Alrabia, M., Alhelewh, N., Al-Salman, A. & Atwell, E. (2014). An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus. *International Journal of Computational Linguistics*, 5(1), 1-13.

Asr, F., Willits, J. & Jones, M. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1092-1097). Philadelphia (Pennsylvania): CogSci.

Baroni, M., Dinu, G. & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238-247). Baltimore: ACL, vol. 1.

Bernier-Colborne, G. & Drouin, P. (2016). Evaluation of distributional semantic models: a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 52-61). Osaka: Computerm.

Bernier-Colborne, G. & L'Homme, M.C. (2015). Using a distributional neighbourhood graph to enrich semantic frames in the field of the environment. In *Proceedings of the Conference Terminology and Artificial Intelligence* (pp. 9-16). Granada (Spain): TIA.

Bertels, A. & Speelman, D. (2014). Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2), 279-303.

Blei, D.M.; Ng, A.Y. & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Bullinaria, J.A. & Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510-526.

Bullinaria, J.A. (2008). Semantic Categorization Using Simple Word Co-occurrence Statistics. In M. Baroni, S. Evert & A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 1-8). Hamburg: ESSLLI.

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). New York: ICML.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. (1990). Indexing by latent semantic analysis". *Journal of the American Society for Information Science*, 41(6), 391-407.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61-74.

El bazi1, I. & Laachfoubi, N. (2016). Arabic Named Entity Recognition using Word Representations. *International Journal of Computer Science and Information Security*, 14(8), 956-965.

Everitt, B., Landau, S. & Leese, M. (2001). *Cluster Analysis* (4th ed.). London: Arnold.

Evert, S. (2007). Corpora and Collocations. Extended Manuscript of Chapter 58 of A. Lüdeling & M. Kytö (Eds.) (2008), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. Retrieved from
http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf
(last access: 2018-12-20).

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, 1, 107-134.

Faber, P. (2011). The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception-action Interface. *Terminology*, 17 (1), 9-29.

Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, P., León-Araúz, P. & Prieto, J.A. (2009). Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies*, 1, 1-23.

Feldman, R. & Sanger, J. (2007). *The Text Mining Handbook*. Cambridge: Cambridge University Press.

Gries, S. & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 73-90). Stanford (California): CSLI.

Hearst, M. & Schütze, H. (1993). Customizing a lexicon to better suit a computational task. In B. Boguraev & J. Pustejovsky (Eds.), *Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics* (pp. 55-69). Columbus (Ohio): ACL SIGLEX.

Herbelot, A. (2015). Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 151-161). London: ACL.

Hermann, M. (2011). Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora. *Journal of Quantitative Linguistics*, 18(1), 23-52.

Jolliffe, I. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Jurafsky, D. & Martin, J. (2017). Vector Semantics. In *Speech and Language Processing*. Draft of August 7, 2017. Retrieved from
https://web.stanford.edu/~jurafsky/slp3/15.pdf (last access: 2018-12-20).

Katrenko, S. & Adriaans, P. (2008). Qualia Structures and their Impact on the Concrete Noun Categorization Task. In M. Baroni, S. Evert & A. Lenci (Eds.), *Proceedings of*

the ESSLLI Workshop on Distributional Lexical Semantics (pp. 17-24). Hamburg: ESSLLI.

Kaufman, L. & Rousseeuw, P. (1990). *Finding Groups in Data*. Hoboken (New Jersey): Wiley-Interscience.

Kazama, J., De Saeger, S., Kuroda, K., Murata, M. & Torisawa, K. (2010). A Bayesian Method for Robust Estimation of Distributional Similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 247-256). Uppsala (Sweden): ACL.

Kiela, D. & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality* (pp. 21-30). Gothenburg (Sweden): EACL.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105-116). Lorient: EURALEX.

Kiss, G. (1973). Grammatical Word Classes: A Learning Process and Its Simulation. *Psychology of Learning and Motivation*, 7.

Krenn, B. (2000). *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*. Saarbrücken: DFKI & Universität des Saarlandes, vol. 7, Saarbrücken Dissertations in Computational Linguistics and Language Technology.

Lapesa, G., Evert, S. & Schulte im Walde, S. (2014). Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics* (pp. 160-170). Dublin: SEM.

León-Araúz, P., Reimerink, A. & Faber, P. (2013). Multidimensional and Multimodal Information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (Eds.), *Computational Linguistics* (pp. 143-161). Berlin: Springer.

León-Araúz, P., San Martín, A. & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations". In *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 73–82). Osaka (Japan): Computerm.

León-Araúz, P., San Martín, A. & Reimerink, A. (2018). The EcoLexicon English Corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the 18th EURALEX International Congress* (pp. 893-901). Ljubljana: Euralex.

Luhn, H. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309-317.

Lund, K., Burges, C. & Atchley, R.A. (1995). Semantic and associative priming in a high-dimensional semantic space. In J.D. Moore & J.F. Lehman (Eds.), *Proceedings*

of the 17th Annual Conference of the Cognitive Science Society (pp. 660-665). Pittsbugh (USA): University of Pittsburgh.

Manning, C.D., Raghavan, P. & Schütze, H. (1998). *Introduction to Information Retrieval*. Cambridge (England): Cambridge University Press.

McInnes, L & Healy, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv e-prints 1802.03426. Retrieved from https://arxiv.org/pdf/1802.03426.pdf (last access: 2018-12-20).

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodogical framework. In D. Bourigault, C. Jacquemin & M.C. L'Homme, (Eds), *Recent Advances in Computational Terminology* (279-302). Amsterdam/Philadelphia: John Benjamins.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*. Scottsdale (Arizona): ICLR.

Miller, G.A. & Charles, W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1-28.

Miller, G.A. (1971). Empirical methods in the study of semantics. In D. Steinberg & L. Jakobovits (Eds.), *Semantics: An Interdisciplinary Reader*. Cambridge: Cambridge University Press; 569–585.

Moisl, H. (2009). Exploratory multivariate analysis. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (vol. 2, pp. 874-899). Berlin: Walter de Gruyter.

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics* (pp. 77-93). Berlin/Munich/Boston: De Gruyter Mouton.

Moisl, H., Maguire, W. & Allen, W. (2006). Phonetic variation in Tyneside: Exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In F. Hinskens (Ed.), *Language Variation – European Perspectives* (pp. 127-141). Amsterdam: John Benjamins.

Moskalski, S. & Torres, R. (2012). Influences of tides, weather, and discharge on suspended sediment concentration. *Continental Shelf Research*, 37, 36-45. Retrieved from https://www.sciencedirect.com/science/article/pii/S0278434312000180 (last access: 2018-12-20).

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3), 291-330.

Nguyen, N.T.H., Soto, A.J., Kontonatsios, G., Batista-Navarro, R. & Ananiadou, S. (2017). Constructing a biodiversity terminological inventory. *PLoS ONE*, 12(4), e0175277.

Pantel, P. & Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining* (pp. 613-619). Edmonton (Canada): KDD-02.

Pennington, J., Socher, R. & Manning, C.D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods for Natural Language Processing* (pp. 1532-1543). Doha (Qatar): EMNLP.

Reimerink, A. & León-Araúz, P. (2017). Predicate-Argument Analysis to Build a Phraseology Module and to Increase Conceptual Relation Expressiveness. In R. Mitkov (Ed.), *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017, Proceedings* (pp. 176-190). London: Springer.

Rohde, D., Gonnerman, L. & Plaut, D. (2006). An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM*, 8, 627-633.

Rojas-Garcia, J., Faber, P. & Batista-Navarro, R. (2018). Conceptual information extraction for named bays from a specialized corpus. In T. Read, S. Montaner & B. Sedano (Eds.), *Technological Innovation for Specialized Linguistic Domains* (pp. 97-113). Beau Bassin (Mauricio): Éditions Universitaires Européenes.

Sahlgren, M. & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 975-980). Austin (Texas): ACL.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1), 33-54.

Salton, G. & Lesk, M.E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1), 8-36.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Stanford (California): Cambridge University Press, Center for the Study of Language and Information Publication, vol. 71, Lecture Notes.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97-124.

Shutova, E., Sun, L. & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp.1002-1010). Beijing (China): COLING, vol. 2.

Spärck J.K., Walker, S. & Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments, part 2. In *Information Processing and Management*, 36, 809-840.

Sun, F., Guo, J., Lan, Y., Xu, J. & Cheng, X. (2015). Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations. In *Proceedings of the 53rd Annual Meeting of the ASL and the 7th International Joint Conference on Natural Language Processing* (pp. 136-145). Beijing (China): ACL.

Sun, L. & Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 638-647). Singapore: EMNLP.

Suzuki, R. & Shimodaira, H. (2004). An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters? In *Proceedings of the Fifteenth International Conference on Genome Informatics*, P034.

Suzuki, R. & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12), 1540-1542.

Thabet, N. (2005). Understanding the thematic structure of the Qur'an: an exploratory multivariate approach. In *Proceedings of the ACL Student Research Workshop* (pp. 7-12). Michigan: ACL.

Thompson, P., Batista-Navarro, R., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M. & Ananiadou, S. (2015). Text Mining the History of Medicine. *PLoS ONE*, 11(1), e0144717.

Turian, J., Ratinov, L., Bengio, Y. & Roth, D. (2009). A preliminary evaluation of word representations for named-entity recognition. In *Proceedings of NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*. Whistler (Canada): Neural Information Processing Systems Foundation.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of Human Language Technology* (pp. 276-283). Edmonton (Canada): HLT/NAACL.

Xu, R. & Wunsch, D. (2009). *Clustering*. Hoboken (New Jersey): IEEE Press/Wiley.

Remote Healthcare Interpreting: a methodology to investigate quality

Esther de Boe[*] - University of Antwerp, Belgium

ABSTRACT

In consonance with our digital era, healthcare interpreting services are increasingly embracing technology. Research on remote public service interpreting indicates that the use of technology adds up to the already complex character of interpreter-mediated communication. Yet, in healthcare settings, little is known about the exact impact of the remote conditions on communication quality. The present explorative study provides a methodological framework to compare two types of remote interpreting (telephone interpreting and video interpreting) with face-to-to face interpreting. Its preliminary findings suggest that there were few differences at the level of message equivalence across the three conditions. However, changes in the interactional dynamics, partly resulting from the remote conditions, did impact on the effectiveness of the communication.
Keywords: remote interpreting; healthcare interpreting; public service interpreting; interpreting quality

RESUMEN

En consonancia con nuestra era digital, los servicios de interpretación sanitaria están adoptando cada vez más la tecnología. La investigación sobre la interpretación a distancia en los servicios públicos indica que el uso de la tecnología se suma al carácter ya complejo de la comunicación mediada por intérpretes. Sin embargo, en los entornos de atención sanitaria, se sabe poco sobre el impacto exacto de las condiciones remotas en la calidad de la comunicación. El presente estudio exploratorio proporciona un marco metodológico para comparar dos tipos de interpretación a distancia (interpretación telefónica e interpretación por vídeo) y la interpretación cara a cara. Sus conclusiones preliminares sugieren que hubo pocas diferencias en el nivel de equivalencia de mensajes entre las tres condiciones. Sin embargo, los cambios en la dinámica interactiva, en parte como resultado de las condiciones remotas, sí afectaron la eficacia de la comunicación.
Palabras claves: interpretación a distancia; interpretación sanitaria; interpretación de servicio público; calidad de la interpretación

* Corresponding author, email: esther.deboe@uantwerpen.be

IN KEEPING WITH our digital era, technology is increasingly used in interpreting services. Especially in the domain of healthcare interpreting, technology is rapidly gaining ground and remote interpreting (RI) by telephone (TI) and video link (VI) is progressively replacing face-to-face (F2F) interpreting. The research in this field is dominated by medical studies that measure the effect of different interpreting modes on equitable access to healthcare, quality of care, efficiency, and patient satisfaction (Pöchhacker and Schlesinger, 2007, p. 4–5). However, empirical research concerning interpreter performance, accuracy, etc. has been largely absent in this setting (Braun and Taylor 2012, p. 45). In other settings, numerous research projects on RI that do empirically examine the interpreting performance, have been undertaken (e.g. Braun and Kohn, 2001; Braun 2003, 2004; Braun and Taylor, 2012; Ko, 2006; Lee, 2007; Oviatt and Cohen, 1992; Rosenberg, 2007; Wadensjö, 1999; Wang, 2017). Although research methods and scopes vary substantially, the studies agree that the use of technology adds up to the already complex character of interpreter-mediated communication.

The present paper reports on the methodological framework which is part of a doctoral research project that was designed to compare quality of interpreter-mediated communication in healthcare settings under three conditions: (1) face-to-face interpreting, (2) telephone interpreting and (3) video interpreting. It attempts to establish to what extent observed problematic issues may be attributed to the remote conditions by investigating relationships between problematic issues at the level of message equivalence, interaction management and technological factors. Moreover, possible benefits and drawbacks of adding image to the audio channel are examined, which is of particular importance to healthcare interpreting, as in this setting, VI is rapidly gaining ground at the expense of TI.

For this purpose, a dataset based on three series of three simulations of interpreter-mediated doctor–patient consultations was created. Within each series of simulations, a professional female interpreter performed her task consecutively in three interpreting modes (F2F, TI, VI). After each simulation, all three participants (doctor, patient and interpreter) were questioned about their perceptions of the quality of the interpreting and the communication, the quality of the provided healthcare and their preferences with regard to the different interpreting modes. The simulations, nine in total, were videotaped, transcribed and annotated. Subsequently, the observed problematic instances were submitted to a comparative, multi-modal analysis, the results of which were triangulated by the findings of the participants' perceptions to establish the influence of the remote conditions on the successfulness of the communication.

The present paper will elaborate on the methodology underlying this research project and will first provide details of the conceptual framework, explaining key concepts, such as quality assessment in interpreting. Subsequently, we will explain the research design, including the different steps taken in the data analysis. Finally, the most salient findings of

the first series of simulations will be described and discussed in the light of previous research on remote interpreting in community settings.

Conceptual Framework

Given that interpreting is a complex cognitive, linguistic, cultural and social process, the issue of quality is multifaceted (Grbić, 2015, p. 334). When assessing the quality of different forms of remote healthcare interpreting as compared to face-to-face interpreting, we must therefore attempt to respect its compound character as much as possible. Following experts in bilingual healthcare communication such as Hsieh (2017) and Meeuwesen et al. (2010), the present study wants to transcend disciplinary-confined approaches to interpreting quality and combine insights from both interpreting studies and studies on healthcare communication. From the perspective of the latter discipline, Hsieh (2017) argues that the interpersonal dynamics between healthcare provider, patient and interpreter may shape the process and content of interpreter-mediated medical encounters. Therefore, if we want to assess the successfulness of intercultural communication in healthcare settings, the examination must incorporate all participants. From a discourse-analytic perspective, adopted by many researchers in interpreting studies, conversations are determined by a process of turn-taking, which is concerned with the way in which participants jointly determine who will speak, who will listen, and how transitions are made between these roles (Goffman, 1981; Gumperz, 1982; Sacks & Jefferson, 1974; Schiffrin, 1994). The participants' collaborative work in constructing turns systematically modifies the structure of each sentence "by adding to it, deleting from it, and changing its meaning" (Goodwin, 1979, p. 112). The interpreter, just like the primary participants, shifts between different roles, expressed through his or her acts of translating, requesting and providing clarifications, engaging in non-verbal behaviour et cetera. Through these actions, the interpreter contributes to the primary participants' interaction, both verbally and non-verbally (Krystallidou, 2014; 2016).

Although in interpreting studies the concept of equivalence (i.e., the relationship between the source text and its rendition in the target language) remains a central issue in the examination of interpreting performances (Pöchhacker, 2004), today, the assessment of equivalence goes far beyond the correct transfer of content. Wadensjö (1998) uses the concept of *closeness* as an indication of source- and target-text correspondence at the linguistic, cultural and interactional level. This correspondence is achieved through accuracy in the transfer of content and by means of coordinating activities (Baraldi & Gavioli, 2014). Therefore, message equivalence and interaction management are closely connected. Of course, measurements of accuracy are extremely complex and should consider not only content, but also style (Hatim & Mason, 1990), including includes affective elements and specific stylistic source message characteristics such as register, which give meaning to the message (Wadensjö, 1998).

Furthermore, in evaluations of interpreting quality, it is important to bear the discourse environment into mind, i.e. the larger, overall institutional and social context in which the interpreter-mediated doctor–patient interaction is embedded (Angelelli, 2004, 26). Just like other institutional communicative events, medical consultations consist of different phases, or genre steps, each of which involves its own specific communication dynamics (e.g., Byrne & Long, 1984; 1976; Heath, 1992; Heritage & Maynard, 2006), and of which the interpreter must be aware (Baraldi & Gavioli, 2014; Tebble, 1993; 1999; 2012).

Independent of which research approach is adopted, we can assume that the purpose of healthcare communication mediated by professional interpreters is to achieve effective communication in terms of mutual understanding between patient and healthcare provider. From the perspective of healthcare communication studies, three aspects that affect the successfulness of the interpreter-mediated interventions are: (1) miscommunication; (2) changes in translation and (3) side-talk activities (Meeuwesen et al., 2010, p. 202). These indicators correspond to the two important dimensions of interpreting quality as considered from the perspective of interpreting studies, discussed earlier in this section, namely (1) message equivalence and (2) interaction management. In the present research, we combine these perspectives on the effectiveness of interpreter-mediated bilingual healthcare communication and assume that instances of miscommunication and changes in translation lead to lower mutual understanding between the primary participants. Therefore, if we want to investigate mutual understanding, we must examine miscommunication. In addition, the communication takes place within a certain interactional environment that impacts of the communication. As a result, we must also look at the way the interaction is managed and examine the actions other than translating (e.g., repair, side-talk activities). In addition, factors independent of the participants, such as issues caused by the environment (e.g., noise) or by the technological conditions in which an interpreter-mediated event takes place (e.g., bad sound quality), may also affect the successfulness of the communication.

Whereas the quality aspects mentioned above can be assessed by means of objective methods, quality can also be assessed via user perceptions. Given that satisfaction with the manner of interpreting and the communication in general is an important key to the establishment of rapport between the participants, examining the perceptions by the users is extremely useful to examine the effectiveness of the communication. Figure 1 represents the most important elements with regard to quality in interpreter-mediated interaction discussed in this section.

Figure 1. Quality of interpreter-mediated communication

Methodology

Research Questions

To evaluate the quality of different forms of remote healthcare interpreting as compared to face-to-face interpreting, we must attempt to respect RI's multifaceted character as much as possible. In line with the focus of this research, the following five research questions were formulated:

1. Are there differences at the level of message equivalence between interpreter-mediated doctor–patient consultations in the three interpreting conditions face-to-face interpreting (F2F), telephone interpreting (TI) and video interpreting (VI)?
2. Are there differences in the interactional dynamics in interpreter-mediated doctor–patient consultations between F2F, TI and VI?
3. Is there a relationship between issues that occur at the level of message equivalence and the interactional dynamics?
4. Can possible quality differences between the remote conditions (TI, VI) and the F2F condition be attributed to environmental and/or technological factors?
5. Are there, within the remote modes (TI, VI), benefits related to the quality of the communication of using an audio-visual channel (VI) in comparison with an audio-only channel (TI)?

Research Design

To respond to these questions, we compared interpreter-mediated interaction in the three conditions. This comparative method, which is frequently used in medical studies on RI (among others Hornberger et al., 1996; Nápoles et al., 2010; Price et al., 2012; Saint Louis et al., 2003), is a useful approach to investigating the impact of the remote conditions on the communication in each condition, while at the same time transcending the idiosyncrasies of the individual interpreters and participants. Despite the artificial nature of role plays, we

decided for this option, to avoid difficulties with obtaining ethics approval for using authentic data, and more importantly, because role plays allow for a systematic comparison. In total, the data set consisted of 9 simulations and 27 post-simulation interviews.

Participants

The three interpreters were based on the following five selection criteria:
1. female interpreter currently working as community/healthcare interpreter;
2. minimum of five years' experience in healthcare interpreting;
3. holder of a university degree;
4. having followed specific interpreting training;
5. member of a professional interpreting organization.

The doctor involved was a practising, experienced gynaecologist who participated on a voluntary basis. The role of the patient was played by a professional actress with ample experience as a simulation patient. Prior to the simulations, informed consent of all participants was obtained.

The language combination that was used was French-Dutch. The doctor was a native speaker of Dutch with non-specialized language proficiency in French; the simulation patient was a native speaker of French with limited knowledge of Dutch.

Procedure

The simulations took place in two rooms. In the doctor's room, the participants were seated in the traditional triadic constellation during face-to-face interpreting. During the remote sessions, the doctor and patient were each seated along a corner of the table, with either the telephone or the tablet in between them. The interpreting room was provided with a telephone and a laptop, equipped with a webcam, while the doctor's room was equipped with a telephone with speaker function and a tablet on a movable display. For the telephone connection, a local landline was used. During the VI sessions, a connection by means of a dedicated platform provided by a professional interpreting agency, which had been arranged and tested well beforehand, had to be abandoned during the implementation of the first series because of technical problems. As a result, Skype video calling had to be resorted to, which consequently had to be repeated for the remaining six simulations of Series II and III, in order to maintain constant experimental conditions. The interpreters were given the choice whether or not to use a headset, which all of them declined, citing reasons such as feeling awkward or unnatural while wearing headsets.

To increase authenticity, the role-plays were designed around three themes from real-life gynaecology practice (Table 1). This was made possible with the help of a second professional gynaecologist. The three main themes were (1) weak pelvic floor; (2) abundant menstruation and (3) pregnancy/infertility. For each themes, three scenarios were created,

which contained slightly different descriptions of the personal circumstances of the patient to avoid repetition for the participants. The scripts contained a minimum of information, to leave space for spontaneous interaction between the participants. The scenarios were played in a different order of conditions, as illustrated in Table 1, to avoid a distorted picture of the exact influence of the remote conditions, since fatigue and cognitive overload may have been heavier by the end of the third simulation. For example, in Series I, during the F2F mode, scenario A of theme 1 (weak pelvic floor) was played, in the VI mode, scenario B of theme 2 (abundant menstruation) was played and in TI, the scenario played was C of theme 3 (pregnancy/infertility).

Table 1. Order of the scenarios across the series and conditions

	Themes/Scenarios		
Series	Theme 1: Weak pelvic floor	Theme 2: Abundant menstruation	Theme 3: Pregnancy/infertility
I	Face-to-face interpreting (F2F): scenario 1A	Video interpreting (VI): scenario 2B	Telephone interpreting (TI): scenario 3C
II	Telephone interpreting (TI): scenario 1B	Face-to-face interpreting (F2F): scenario 2C	Video interpreting (VI): scenario 3A
III	Video interpreting (VI): scenario 1C	Telephone interpreting (TI):	Face-to-face interpreting (F2F): scenario 3B

Data Analysis

Annotation

As a first step in the process of analysing the effectiveness of the communication, the transcriptions of the video-recorded simulations were annotated. For this purpose, an annotation scheme was designed (Table 2), which combines the most important aspects related to quality as discussed in the section 'Conceptual Framework', as well as contextual factors, which are of specific interest to examine the impact of the remote conditions. This resulted in the creation of three main annotation categories:

1. Issues (message equivalence and interaction)
2. Interaction management (repair and optimization of the communication and backchannelling)
3. Environment & technology

Table 2. Annotation model

(1) Issues		(2) Interaction management		(3) Environment & technology
Message Equivalence	**Interaction**	**Repair & Optimization**	**Backchannelling**	
Omission	Overlapping speech	Request for clarification	Reactive expression	Operation of equipment
Addition	False start	Clarification/ Repetition	Resumptive opener	Image quality
Substitution	Pause	Meta comment	Gaze/Mutual gaze	Sound quality
False fluency	Near false start	Correction	Posture	Synchronisation
	Language mixing	Compensation	Gesture	Out of shot
	Reported speech	Transform reported speech	Facial expression	Seating
		Renditional formulation	Nod/Shake head	Noise
		Change in order	Repetition	External disturbance

The annotation model is based on the assumption that there are two parallel levels of communication. The first level is the main level of communication, which is the linguistic content as expressed by the participants, wrapped up in the paralinguistic features (e.g., register, prosody) that give meaning to it. When the source-text content, including the paralinguistic features, is altered or omitted in the rendition in the target-text by the interpreter, this is considered an issue at the level of message equivalence. To analyse shifts in message equivalence, a combination of categories used to express accuracy in interpreting studies and medical studies based on Braun & Taylor (2012), Balogh & Hertog (2012) and Flores (2012) was applied. In total, four categories of issues were annotated: (1) omission: the interpreter omits relevant information or affective elements; (2) substitution: the interpreter alters information or affective elements; (3) addition: the interpreter adds information or affective elements, and (4) false fluency: the interpreter uses a word/phrase that does not exist in the target language or an incorrect word/phrase that substantially alters its meaning (Flores et al., 2012, p. 546).

Parallel to the main level, communication takes place at an interactional level. This level provides a context to the way information is exchanged at the main level. During the interaction between the participants, issues which inhibit smooth communication may occur. These interactional issues can become problematic in interpreter-mediated communication and may be heightened in remote interpreting (Braun & Taylor, 2012). Issues which were annotated in this category were overlapping speech, often in combination with or caused by pauses and/or false starts, as well as near false starts. Other interactional issues which may indicate underlying communication problems and/or cognitive overload, are reported speech and language mixing. In the case of reported speech, the use of the personal pronoun, a form to be used preferably by interpreters according to most codes of ethics, is abandoned and a participant is referred to by means of the addition of a reporting verb (e.g., "he says") and a change in personal pronoun (e.g., from "I" into "he" or "she") (Bot, 2005). In the case of language mixing, the interpreter speaks in two languages within a single turn, which can happen due to confusion and/or cognitive overload. The interactional phenomena of reported speech and language mixing can also be part of repair and optimization strategies.

At the interactional level, all participants are engaged in the conversation management by giving each other feedback on the comprehension of the information exchanged at the main level. They do this in order to optimize the communication and prevent miscommunication, for example by explicitly commenting on the discourse meta structure (e.g., when the doctor announces: "I will now ask you a few short questions, one by one"). Interaction management can also be aimed at repairing miscommunication, for example by explicitly requesting clarification. Other repair and optimization strategies that were annotated include clarifications/repetitions, meta comments, corrections, compensations, changes in order and renditional formulations. Renditional formulations are interpreter renditions specifically designed to summarise, gloss or develop the gist of the previous answer (Baraldi & Gavioli, 2012; Heritage, 1985).

The signalling of non-comprehension or non-agreement with an utterance by another speaker can also be expressed in a less explicit way, such as facial expressions. We will refer to these signals as 'backchannelling', i.e. visual behaviour by means of which speakers and listeners derive key information to clarify content and provide feedback (Whittaker & O'Connaill 1997, p. 28–29): gaze, gesture, posture and facial expression.

Together, the main level and the interactional level form a subtle verbal and nonverbal interplay between listener and hearer, who constantly exchange turns following a certain turn-taking pattern, in line with their discourse environment. The two levels can also affect each other, and can in their turn be directly or indirectly influenced by external factors that impact on smooth communication, due to the environment or technological conditions. This annotation category consists of problems with the operation of equipment, image quality, sound quality, synchronization of image and sound, out of shot (being within the field of the

camera or not), seating (the way the participants are seated within the room), noise and external disturbances.

Identification of Issues

After having been annotated, all issues at the level of message equivalence and interaction were submitted to a quantitative analysis. It must be stressed that, given the small size of the corpus, the purpose of this was not to generate statistically significant data, but to identify problematic issues and subsequently enable a qualitative examination of relationships between issues occurring at different communication levels at the same time.

At the level of message equivalence, all annotated issues were scrutinized and ranked as *problematic* issues, if they led to a loss of information or change of meaning, with possible clinical consequences (Flores, 2012) or a loss of empathy and emotions. The reasoning behind this was that, if all typical error categories, such as omissions, were considered inaccuracies at the level of single turns or adjacency pairs (question–answer pairs), the complex interactional process of achieving mutual understanding through verbal and nonverbal resources and by coordinating activities would be overlooked. However, it must be noted that the assessment of the problematic character of the annotated issues was not always clear-cut and involved a certain degree of subjectivity, especially since there was only one annotator, given the limited size of the present research project. In cases of doubt about whether a certain issue was to be considered problematic or not, the video recordings of the moment in which a particular instance happened were re-examined thoroughly, as well as the turns preceding and following this instance.

Excerpt 1 illustrates the difference between a message equivalence issue and a *problematic* message equivalence issue: in the F2F session of Series I, the Interpreter (I) omits (OM, turn 79) a part of the instructions provided by the Doctor (D) in turn 78, namely the fact that the Patient (P) has to hand in urine at the laboratory. Since the treatment had been discussed earlier on during the consultation, there is no clear risk of misunderstanding. However, in the same rendition, a reference to patient participation is substituted (SUB: "to discuss the results together" → "we are going to see how everything is going"), which is considered as problematic. Both issues are marked in bold in Excerpt 1.

Excerpt 1. Example of problematic and non-problematic ME issues

79	D: […] daar geef ik de papieren voor mee, **dan kunt u bij het laboratorium urine inleveren,** krijgt u bij de balie de vragenlijsten, verwijs ik u naar de fysiotherapeut en als u daar de intake hebt gehad en dat loopt, komt u nog een keer terug **om de uitslag samen te bespreken.**	D: […] for that I will give you the papers **then you can hand in urine at the laboratory,** you will receive the questionnaires at the desk, I direct you towards the physical therapist and when you will have had your intake there and it's running, you will return once more to me **to discuss the results together.**
80 O M SU B	I: Euh, je vais vous donner un papier pour **aller au laboratoire**. Vous pouvez chercher à l'accueil également le questionnaire pour le compléter. Je vais prescrire la physiothérapie. Et puis je vais vous demander après un certain temps de revenir vers moi, **on va voir comment tout se passe** […].	I: Euh, I am going to give you a paper **to go to the laboratory.** You can also go to the desk for the questionnaire to complete it. I am going to subscribe the physiotherapy. And also I am also going to ask you after a certain time to come back to me, **we are going to see how everything is going** […].

Note: D = Doctor, I = Interpreter, SUB = substitution, OM = omission
For a further explanation of transcription and translation decisions, see endnote [1]

Apart from the identification of issues, an inventory was made of the interaction management, i.e. strategies aimed at repairing miscommunication and optimizing communication. These quantitative results were compared across the three conditions (face-to-face, telephone and video interpreting).

Analysis

Both at the level of message equivalence and at the level of interaction, a distinction was made between issues which concurred with issues in other categories (concurrences) and issues which happened independently from issues in other categories (non-concurrences). Subsequently, in order to establish the role of environment and technology in miscommunication, observed instances of concurring problematic issues were submitted to multi-modal analyses. In the multi-modal analyses, the way the participants dealt with the interaction management was also taken into account. Finally, to evaluate the subjective effectiveness of the communication, the qualitative analyses were triangulated by the findings from the participants' perceptions.

Preliminary Findings

In this section, we will summarize the preliminary findings of the analyses of the first series of simulations. Consecutively, we will discuss issues, the concurrence of issues at different levels and the interaction management. Examples from the data analysis will be provided to illustrate differences in the interactional dynamic between the three conditions.

Issues

Since there were no written-out scenarios, the three simulations varied in their total duration. To be able to compare the relative weight of the issues which occurred in the three role plays, we synchronized them on the basis of their average duration in seconds. The relative weight of the number of problematic issues that happened (Table 3) was calculated as follows: the absolute number of issues was multiplied by the average duration of all nine simulations (1250 seconds). Subsequently, the resulting product was divided by the total duration of the simulation in question in seconds. In Table 3, we read, for example, that in the face-to-face session of Series I, the absolute number of problematic message equivalence issues which occurred was 33. The total duration of this simulation was 23 minutes and 59 seconds, or 1439 seconds, while the average duration of the simulations was being 1250 seconds. This was calculated as: 33 (absolute number of issues) x1250 (average duration in seconds of the simulations/1439 (duration in seconds of the simulation under analysis) = 28.66. In a last step, this outcome was rounded up to 28.7, which is the relative number of issues in the face-to-face session of Series I.

Table 3. Quantative summary issues Series I

	Interpreting condition								
	Face-to-face interpreting (F2F)			Telephone interpreting (TI)			Video interpreting (VI)		
Issues	ABS	REL	CONC	ABS	REL	CONC	ABS	REL	CONC
ME	33	28.7	30.3%	37	26.1	32.4%	32	26.7	50.0%
IA	12	10.4	58.3%	22	15.5	50.0%	21	17.8	80.9%

Note. ME= message equivalence, IA= interaction, ABS/REL= absolute/relative incidence, CONC = concurrence

As far as miscommunication at the level of message equivalence is concerned, the analyses showed few quantitative differences across the conditions (F2F: 28.7/TI: 26.1/VI: 26.7). The reason why in, in spite of the increased relationship between the issues at different levels during remote interpreting, the number of issues at the level of message equivalence was finally comparable across the conditions, may be linked to differences in the average turn duration. Since the turns were much longer in the F2F session (24.81 seconds), almost twice as long as the average duration in TI (15.82 seconds) and VI (14.92 seconds), the cognitive strain on the interpreter may have been higher during the F2F session, resulting in an increase in the number of equivalence issues. Contrary to the level of message equivalence, at the interactional level, differences between the remote modes (TI: 15.5/VI: 17.8) and the F2F mode (10.4) can be observed, with the highest number of issues occurring in the VI mode, although the difference in number between the two remote modes is relatively small.

With regard to concurrence, both at the level of message equivalence and interaction, the F2F and TI mode show similar percentages (F2F-ME: 30.3%; F2F-IA: 58.3%; TI-ME:

32.4%; TI-IA: 50%), whereas in the VI mode, at both levels, concurrence was clearly higher (VI-ME: 50.0%; VI-IA: 80.9%). Of course, these numbers are only meant to identify possible causes of miscommunication and potential relationships between issues concurring at different levels, which were further explored by means of multimodal analyses.

Interaction Management
Compared with the F2F condition (12.2 instances of repair and optimization), the interaction management was increased in the TI condition (24.3), and even more so in the VI session (38.3).

Table 4. Quantative summary repair and optimization Series I

	Interpreting condition					
	Face-to-face interpreting (F2F)		Telephone interpreting (TI)		Video interpreting (VI)	
	ABS	REL	ABS	REL	ABS	REL
RO	14	12.2	28	24.3	46	38.3

Note. RO = repair and optimization, ABS/REL = absolute/relative incidence

In the next section, we will illustrate the differences in the way miscommunication occurred across the conditions by means of few examples, concerning the most prevailing interactional issue, namely overlapping speech. We will also discuss differences in the way the interaction was managed across the conditions.

Face-to-face interpreting
Although overlapping speech dominated the interactional issues in all conditions, its effect differed in the remote modes in comparison with the face-to-face mode. During F2F, overlapping speech happened frequently, but concurred only in a few cases with issues at other levels. When it occurred, non-verbal and verbal backchannelling between the participants sufficed to quickly repair the miscommunication. This is illustrated in Excerpt 2: the Interpreter leaves a short pause after her rendition and then corrects herself (COR) during the same turn (39) by adding "not to be pregnant". Meanwhile, the Patient has understood the message and has started answering "I did not start again with the pill", causing a false start (FS, turn 40), leading to overlapping speech (OS). The Interpreter requests clarification (RCL, turn 41) by asking "After?", directly to the Patient. The Patient repeats her phrase and adds "after": "I did not start again after" (CLA, turn 42). The Interpreter summarizes these turns in one renditional formulation (RF, turn 43). The only content that is lost is: "for the moment", which is not considered crucial for understanding the message or adapting the message in an important way.

Excerpt 2. (F2F-1)

39 **PS CORR**	I: qu'est-ce que vous utilisez pour protéger (..) **pour ne pas être enceinte?**	I: what do you do to protect? (..) not to be **pregnant**?
40 **FS→OS**	P: **Pour l'instant** (.) on utilise des préservatifs, j'ai pas repris la pilule.	P: **For the moment** (.) we are using condoms, I did not start again with the pill.
41 **RCL**	I: Après?	I: After?
42 **CLA**	P: J'ai pas repris après.	P: I didn't start again after.
43 **RF**	I: Ik ben niet herbegonnen met mijn pil, wij gebruiken enkel preservatieven.	I: I have not restarted the pill, we are only using condoms.

Note. PS = pause, CORR = correction, FS = False start, OS = overlapping speech, RCL = request for clarification, CLA = clarification, RF = Renditional formulation
(.) = pause of 1 second, **bold** = overlapping speech

Telephone Interpreting

In the TI mode, we observed that repair of overlapping speech was more difficult than in the face-to-face mode. This is illustrated in Excerpt 3: the Interpreter has trouble comprehending the Doctor's message (turn 41), which is expressed by her frowning and hesitation. There is a problematic substitution (SUB) in her rendition (42) of the Doctor's turn, that changes the message content: important information is missing, namely the burden the treatment may put on the couple, while the idea that both the Patient and her husband have a chance to succeed is added. Moreover, the order in which information was given is changed (CO): the Interpreter places the chances of succeeding "per cycle" at the end of the rendition, after a short, intra-turn pause. Meanwhile, the Patient has already started asking for clarification (RCL, turn 43), causing overlapping speech (OS) which in its turn leads to a series of repair strategies (turns 44–47). What adds up to the miscommunication, is noise in the hallway. Since the volume transmitted by the speaker mode was already reduced and caused a slight echo, noise from outside complicated repair strategies and required more efforts from the participants to understand what was being said. Both Doctor and Patient lean ostentatiously forward to listen. At the same time, the Interpreter corrects herself twice during her rendition (COR), rendering the communication flow altogether less smooth.

Excerpt 3. (TI-1)

41	D: […] Euh, en ik kan zo meteen uitleggen wat een cyclus van ivf euh inhoudt euhm en de belasting voor u en voor uw partner, de kans per cyclus dat u zwanger wordt is ongeveer 25 procent.	D: […] Euh, and I can explain you in a short while what a ivf cycle means euhm and the strain it puts on you and your partner, the chance per cycle that you get pregnant is around 25 per cent.
42 **SUB** **COR** **CO**	I: […]Euhm, moi je vais vous expliquer qu'est-ce qu'un cycle de in vitro fait, qu'est-ce que ça signifie et les chances que vous avez alors de de réussir et c'est pour vous et pour votre partenaire de 25 pour cent (..) **Par cycle.**	I: […]Euhm, I am going to explain to you what an in vitro cycle does what it means and the chances you have well to succeed and it's for you you and your partner 25 per cent (..) **Per cycle.**
43 **OS** **RCL**	P: **25 pour cent de chance** de plus d'avoir un enfant?	P: **25 per cent of chance** more to have a child?
44 **COR**	I: 20...25 per cent kans meer om zwanger te worden (.) om een kind te krijgen?	I: 20...25 per cent chance of getting pregnant (.) to have a child?
45 **RCL**	P: Ou bien dans 25 pour cent des cas je tombe enceinte?	P: Or in 25 per cent of the cases I get pregnant?
46 **COR**	I : Of is het dat ik in 5..25 per cent gevallen zwanger word?	I: Or is it that in 5..25 per cent of the cases I get pregnant?

Note. SUB= substitution, COR= correction, CO = change of order, OS= overlapping speech, RCL= request for   clarification
(..) = pause of 2 seconds, **bold** = overlapping speech

Video Interpreting

Of the three conditions, instances of overlapping speech were most problematic during the VI mode. The increased interaction management resulted partly from the interactional behaviour by the primary participants: spontaneous reactions, for example, when the Doctor did not wait for the rendition by the Interpreter when she had understood the Patient, especially during the short verbal examination (involving yes/no answers), caused overlapping speech. In the VI mode, overlapping speech caused a distortion of the sound and often resulted in disruption of the communication flow. In addition, repair was more complicated and often, chain reactions of issues happened: an issue at one level caused issues at other levels, while during the repair of the miscommunication, additional issues occurred which in their turn had to be repaired. An example of such a chain reaction in provided in Excerpt 4: the Doctor has asked the Patient what kind of pain killer she is using. The Patient is trying to find the name of the medicine and is speaking hesitantly, with

reduced voice quality (turn 34). When the Patient pauses, the Interpreter overlaps (OS) by requesting clarification (RCL) to the Patient (turn 35). The Doctor helps to solve the problem by asking the Patient if it is a kind of paracetamol (RCL, turn 36), which the Patient acknowledges (CLA, turn 37), in overlap with the Doctor's answer (OS). This causes no problems of understanding between the Doctor and the Patient, but since the overlapping speech disturbs the sound quality, the Patient's answer is inaudible to the Interpreter, who provides a renditional formulation (RF, turn 40) in reported speech (RS), summarizing the previous turns. In fact, this rendition is redundant, since the Doctor and Patient have already reached mutual understanding. However, since the Interpreter does not have full access to all the backchannelling between the Doctor and the Patient in turn 36–37, she cannot follow the communication. Her frowning, smiling and constant manipulation of the mouse point to stress and insecurity. Moreover, the Interpreter's strategy of translating everything that has been said, including meta communication (obliged by most interpreters' codes of ethics), makes the communication less efficient, while the reported speech increases the distance between the participants.

Excerpt 4 (VI-1)

33	P: Perdo feminin ou Perdolan euh (.) bon euh oui (.) **Ça dépend.**	P: Perdo feminin or Perdolan euh (.) well euh yes (.) **It depends.**
34 OS RCL	I: **Euh, vous avez** dit Perdolan seulement?	I: **Euh, did you** only say Perdolan?
35 CLA	P: Perdo feminin ou Perdolan euh d'antidouleur comme ça euh **(.)** du euhm, comment on l'appelle encore (..) oui, pas de, oui enfin, tout ce que je trouve d'antidouleur.	P: Perdo feminin or Perdolan euh pain killers like that **(.)** of euhm, how are they called again (..) yes, not, yes well actually, any pain killer I find.
36 RCL	D**: Is het een vorm van paracetamol?**	D: **Is it a kind of paracetamol?**
37 OS CLA	P: **Oui, voilà.**	P: **Yes, that's it.**
38 META MIX	I: **Oui, euh, het…, hebt u het begrepen?**	I: **Yes euh, [*Dutch*] it…, did you understand?**
39 OS	P: **C'est du paracétamol.**	P: **It's paracetamol.**

40	I: Paracetamol. Mevrouw zei Perdolan of	I: Paracetamol. Madam said Perdolan or
REP	Perdolan voor vrouwen, Perdo féminin,	Perdolan for women, Perdo féminin, euhm
RF	euhm alles wat ik kan gebruiken zei ze dan,	anything I can use, she said then, yes, of pain
RS	ja, aan pijnstillers.	killers.

Note. OS= overlapping speech, RCL= request for clarification, CLA= clarification, META-MIX: meta comment-language mixing, REP= repetition, RF= renditional formulation, RS= reported speech

Conclusion and Discussion

This paper has presented a methodology for investigating possible differences in the quality of interpreter-mediated communication by telephone and video link as compared to face-to-face interpreting. In addition, we have discussed the preliminary results of the first series of simulations, which are part of a larger research framework consisting in total of nine simulations and involving three different interpreters.

When we evaluate the effectiveness of the communication in the first three simulations, we can conclude the following. As far as miscommunication at the level of message equivalence is concerned, the quality assessment showed few irregularities between the conditions. This corroborates findings by Crossman et al. (2010), who verified the concordance between healthcare providers original diagnosis and its rendition in face-to-face and remote interpreting, and observed no significant differences. In addition, the types of issues were similar across the conditions: more than half of the problematic equivalence issues were omissions.

At the interactional level, we did observe an increased number of interactional issues during both remote modes. Concerning the types of interactional issues, however, there were no differences between the conditions, with overlapping speech dominating in all three conditions. In the remote modes, overlapping speech concurred more often with issues at the level of message equivalence and technological and environmental issues than during F2F. This is in line with findings in VI research by Balogh & Hertog (2012), Balogh & Salaets (2018) and Braun (2004, 2007), who reported an increase in interactional issues and established a correlation between overlapping speech and omission in remote conditions.

Not surprisingly therefore, the assessment of the interaction management showed large differences in the number of repair and optimization strategies between F2F, TI and VI, with the highest number of repair and optimization strategies occurring in VI. This confirms that in the remote conditions, an increased amount of the communication that was attributed to interaction management, as was found in previous research on TI (e.g., Oviatt & Cohen, 1992; Rosenberg (2007; Wadensjö, 1999) and VI (e.g., Braun 2015).

The increased number of interactional issues, the heavier interaction management and elevated number of strategies of repair and interaction in the technology-mediated simulations, made the communication less successful in the remote conditions. Moreoever,

possible positive benefits of adding video and thus being able to see one another (as opposed to the TI sessions) were in fact undermined by the effect of the technical limitations of the video conferencing equipment, as a result of which overlapping speech caused sound quality problems, creating a need for repair and complicating repair at the same time.

Of course, given the small scale, these findings are not representative and must be considered as a case study. Moreover, the findings are preliminary and need to put in the perspective of the larger research framework. However, the analyses of the these first three simulations shows that the designed methodology is well-suited to study the multidisciplinary phenomenon of remote interpreting in a systematic way, and already point at interesting differences in interactional dynamics between the three conditions. The analysis of the three series together should provide us with a clear picture of the shifts in interactional dynamics which we may expect in technology-mediated healthcare interpreting, and of the role of technology in this. This knowledge will be useful to all practioners of interpreter-mediated remote communication in healthcare: doctors, patients and interpreters.

References

Angelelli, C.V. (2004). *Medical Interpreting en cross-cultural communication*. Cambridge: Cambridge University Press.

Baker, M. (2006). Contextualization in translator- and interpreter-mediated events. *Journal of Pragmatics 38*, 321–337.

Bot, H. (2007). Dialogue interpreting as a specific case of reported speech. In F. Pöchhacker and M. Schlesinger (Eds.) *Healthcare Interpreting: Discourse and Interaction*. Amsterdam/Philadelphia: John Benjamins.

Balogh, K. & Hertog, E. (2012). In S. Braun & J. Taylor (Eds.). *Videoconference and remote interpreting in criminal proceedings*. (pp. 101–116). Guildford: University of Surrey. Retrieved from http://www.videoconference-interpreting.net/BraunTaylor2011.html.

Balogh, K. & Salaets, H. (2018). Videoconferencing in legal context: A comparative study of simulated and real-life settings. In J. Napier, R. Skinner and S. Braun (eds.). *Here or there. Research on interpreting via video link*. (pp. 264–296). Washington: Gallaudet University Press.

Baraldi, C. & Gavioli, L. (2012). *Coordinating participation in dialogue interpreting*. Amsterdam/Philadelphia: John Benjamins.

Baraldi, C. & Gavioli, L. (2014). Are close renditions the golden standard? Some thoughts on translating accurately in healthcare interpreter-mediated interaction. *The Interpreter and Translator Trainer 8*(3), 336–353. Retrieved 4 May 2016 from https://doi.org/10.1080/1750399X.2014.972029.

Bot (2005). Dialogue Interpreting as a specific case of reported speech. *Interpreting 7*(2), 237–261.

Braun, S. (2003). Kommunikation unter widrigen Umständen? - Optimierungsstrategien in zweisprachigen Videokonferenz-Gesprächen. In J. Döring, W.H. Schmitz, O.A. Schulte (Eds.), *Connecting Perspectives. Videokonferenz: Beiträge zu ihrer Erforschung und Anwendung* (pp. 167–185). Aachen: Shaker.

Braun, S. (2004). *Kommunikation unter widrigen Umständen? Fallstudien zu einsprachigen und gedolmetschten Videokonferenzen*. Tübingen: Narr.

Braun, S. & Taylor, J. (2012). *Videoconference and remote interpreting in criminal proceedings*. Guildford: University of Surrey.

Braun, S. & Kohn, K. (2001). Dolmetschen in der Videokonferenz. Kommunikative Kompetenz und Monitoringstrategien. In G. Thome, C. Giehl and H. Gerzymisch-Arbogast (Eds.), *Kultur und Translation: Methodologische Probleme des Kulturtransfers* (pp. 3–32). Tübingen: Narr.

Byrne, P.S. & Long, B.E.L. (1984/1976). *Doctors talking to patients: A study of the verbal behaviours of doctors in the consultation*. Exeter: Royal College of General Practitioners.

Flores, G. et al. (2012). Errors of medical interpretation and their potential clinical consequences: a comparison of professional versus ad hoc versus no interpreters. *Annals of Emergency Medicine 60*(5), 545–553.

Goffman, E. (1981). *Forms of Talk*. Philadelphia: Pennsylvania Press.

Grbić, N. (2015). Quality. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (pp. 333–336). London and New York: Routledge.

Goodwin, C. (1979). "The Interactive Construction of a Sentence in Natural Conversation". In G. Psathas (Ed.), *Everyday Language: Studies in Ethnomethodology*. (pp. 97–121). New York: Irvington Publishers.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.

Hale, S. (2004). *The Discourse of Court Interpreting. Discourse practices of the law, the witness and the interpreter*. Amsterdam/Philadelphia: John Benjamins.

Heath, C. (1992). The delivery and reception of diagnosis in the general-practice consultation. In P. Drew and J. Heritage (Eds.), *Talk at work: interaction in institutional settings* (pp. 235–267). Cambridge: Cambridge University Press.

Heritage, J. & Maynard, D.W. (2006). *Communication in Medical Care: Interaction between primary care physicians and patients*. Cambridge: Cambridge University Press.

Hornberger, J. et al. (1996). Eliminating language barriers for non-English-speaking patients. *Medical Care* 34 (8), 845–856. Retrieved 1 April 2017 from http://www.ncbi.nlm.nih.gov/pubmed/8709665

Hsieh, E. (2017). The Model of Bilingual Health Communication: Theorizing interpreter-mediated medical encounters. In E.A. Jacobs & L.C. Diamond (Eds.) in *Providing*

health care in the context of language barriers: International perspectives (pp. 35–55). Bristol, UK: Multilingual Matters.

Ko, L. (2006). The need for long-term empirical studies in remote interpreting research: a case study of telephone interpreting. In E. Hertog & B. van der Veer (Eds.). *Linguistica Antverpiensia 5*, 325–338.

Krystallidou, D. (2014). Gaze and body orientation as an apparatus for patient inclusion into/exclusion from a patient-centred framework of communication. *The Interpreter and Translator Trainer 8*(3), 399–417.

Krystallidou, D. (2016). Investigating the interpreter's role(s). The A.R.T. framework. *Interpreting 18*(2), 172–197.

Lee, J. (2007). Telephone Interpreting – seen from the interpreter's perspective. *Interpreting 9*(2), 231–252.

Meeuwesen, L. Twilt, S., Ten Thije, J.D & Harmsen, H. (2010). 'Ne diyor?'' (What does she say?): Informal interpreting in general practice. *Patient Education and Counseling* 81, 198–203.

Merlini, R. and Favaron, R. (2007). Examining the voice of interpreting in speech pathology. In F. Pöchhacker & M. Shlesinger (Eds.), *Healthcare Interpreting* (pp. 101–137). Amsterdam/Philadelphia: John Benjamins.

Napier, J. (2004). Interpreting Omissions: A new perspective. *Interpreting 6*(2), 117–142.

Napoles, A.M., Santoyo-Olsson, J., Karliner, L.S., O'Brien, H., Gregorich, S.E., Pérez-Stable, E.J. (2010). Clinician ratings of interpreter mediated visits in underserved primary care settings with ad hoc, in-person professional, and video conferencing modes. *Journal of health care for the poor and underserved 21*, 301–17. Retrieved 1 May 2016 from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576468/.

Oviatt, S. & Cohen, P.R. (1992). Spoken language in interpreted telephone dialogues. *Computer Speech and Language 6*, 277–302.

Pöchhacker, F. (2004). Introducing Interpreting Studies. London/New York: Routledge.

Pöchhacker, F. & Schlesinger, M. (Eds.). (2007). *Healthcare Interpreting*. Amsterdam/Philadelphia: John Benjamins.

Price Leemann, E. et al. (2012). Interpreter perspectives of in-person, telephonic, and videoconferencing medical interpretation in clinical encounters. In *Patient Education and Counseling 87*(2), 226–232.

Rosenberg, B.A. (2007). A data driven analysis of telephone interpreting. In C. Wadensjö, B. Englund Dimitrova & A.L. Nilsson (Eds.), *The Critical Link 4. Professionalization of interpreting in the community* (pp. 65–76). Amsterdam/Philadelphia: Benjamins.

Sacks, H. & Jefferson, G. (1974). "A simplest systematics for the organization of turn-taking in conversation". *Language 50*, 696–736.

Schiffrin, D. (1994). *Approaches to discourse*. Cambridge, MA/Oxford: Blackwell.

Saint-Louis, L. et al. (2003). *Testing new technologies in medical interpreting*. Somerville, Massachusetts: Cambridge Health Alliance. Retrieved 27 November 2016 from http://www.challiance.org/Resource.ashx?sn=CommunityAffairstnthndbk.

Tebble, H. (1993). A discourse model for dialogue interpreting in Australian Institute of Interpreters and Translators (AUSIT) in *Proceedings of the first practitioners' seminar*, Canberra: NAATI.

Tebble, H. (1999). The tenor of Consultant Physicians. Implications for medical interpreting. *The Translator 5*, 179–200.

Tebble, H. (2012). Interpreting of interfering? In C. Baraldi and L. Gavioli (Eds.), *Coordinating participation in dialogue interpreting* (pp. 23–44). Amsterdam/Philadelphia: John Benjamins.

Wadensjö, C. (1998). *Interpreting as Interaction*. London/New York: Longman.

Wadensjö, C. (1999). Telephone Interpreting & the Synchronization of Talk in Social Interaction. *The Translator 5*(2), 247–264.

Wang, J. (2017). 'Telephone interpreting should be used only as a last resort.' Interpreters' perceptions of the suitability, remuneration and quality of telephone interpreting". *Perspectives: Studies in Translation Theory and Practice*. doi 10.1080/0907676X.2017.1321025.

Whittaker, S. & O'Connaill, B. (1997). The role of vision in face-to-face and mediated communication". In K.E. Finn et al. (Eds.), *Video-mediated communication* (pp. 23–49). Mahwah: Lawrence Erlbaum Publishers.

Endnote

For the transcriptions, created our own conventions instead of using more detailed ones such as the much used transcription conventions proposed by Jefferson (1974). This choice was guided by the holistic approach of the present research, which wants to examine several phenomena at the same time (equivalence issues, interaction issues, and the influence of environmental and technological factors) and illustrate these issues by means of examples. In these examples, a granular description of all discourse-analytic features is not necessarily relevant and may rather inhibit the readability of the represented exchanges. With respect to the translations of the transcriptions (which are in Dutch and French, the languages used in the simulations) in English, we have opted for a pragmatic approach (Hale 2004: 213), and have attempted to remain close to the source text, while at the same time avoiding ungrammatical constructions or non-idiomatic expressions, unless this was the case in the original utterance or rendition.

Multilingualism 2.0: language policies and the use of online translation tools on global platforms

Irene Cenni[*] - Ghent University, Belgium

ABSTRACT

At present, the existing 2.0 Web is far more multilingual than was ever anticipated in the early days of the Internet (Hale, 2014; Hale & Eleta, 2017). Indeed, the increasing variety of languages is a phenomenon that signals the end of the first stages of the digital era in which the Internet was characterized by English-language dominance (Leppänen & Peuronen, 2012). This study primarily aims to present the emerging topics in multilingual research that focus on 2.0 platforms. It presents a literature review and discusses a number of multilingual strategies adopted by different platforms. Five popular platforms have been considered, namely Wikipedia, Facebook, Instagram, Booking.com and TripAdvisor, with close attention paid to travel platforms (and TripAdvisor, in particular). For 2.0 platform providers such as TripAdvisor, multilingualism constitutes a challenge. Typically, these platforms do not opt for an English-only rule, but rather develop linguistic policies in order to accommodate their multilingual users (Cenni & Goethals, 2017). The case of TripAdvisor is particularly striking, not least because it is characterized by the coexistence of two divergent multilingual strategies on the same platform.
Keywords: multilingualism online, 2.0 (travel) platforms, 2.0 platform design, online machine translation, multilingual user-generated content

RESUMEN

En la actualidad, la Web 2.0 tiene un carácter mucho más multilingüe de lo que se anticipó a principios de la era de Internet (Hale, 2014; Hale & Eleta, 2017). De hecho, el progresivo uso de diferentes lenguas es un fenómeno que continúa creciendo, dejando atrás las primeras etapas de la era digital en las que Internet se destacaba por el dominio sin igual del inglés (Leppänen & Peuronen, 2012). El presente estudio tiene como finalidad presentar aquellos temas que emergen en la investigación multilingüe centrada en las plataformas 2.0. Se presenta una revisión bibliográfica, junto con una discusión acerca de diferentes estrategias adoptadas por diversas plataformas. Se analizan cinco plataformas, en concreto Wikipedia, Facebook, Instagram, Booking.com y TripAdvisor. El estudio presta especial atención a las plataformas de viajes, TripAdvisor en particular. Normalmente, estas plataformas no optan por el uso exclusivo del inglés, sino que desarrollan políticas lingüísticas orientadas a incluir a los usuarios multilingües (Cenni & Goethals, 2017). El caso de TripAdvisor es particularmente notable al caracterizarse por dos estrategias multilingües divergentes dentro de la misma plataforma.
Palabras clave: multilingüismo online, plataformas de viaje 2.0, diseño de plataformas 2.0, traducción automatizada online, contenido multilingüe generado a partir de los usuarios

* Corresponding author e-mail: irene.cenni@ugent.be

INTERNET USERS ARE reading and writing in multiple languages as never before, and the percentage of online content written in English is in stable decline (Hale, 2012; 2014). This picture was quite different in the early days of the Internet. During the 1990s, English was the unrivalled language adopted in the digital context, mainly because it represented the language of the vast majority of website developers and users, and English was the language of interaction among speakers of different linguistic backgrounds (Danet & Herring, 2007; Leppänen & Peuronen, 2012). At that time, the linguistic landscape could easily be described as primarily monolingual, with English as the dominant language (Leppänen & Peuronen, 2012). Since then, the range of languages used on the Internet has rapidly and dramatically changed. For instance, other languages like Chinese, the second most-used language online, could challenge the status of English online in a not too far future (Kelly-Holmes & Lenihan, 2017).

Undoubtedly, the online environment has grown into a linguistically plural setting and multilingualism online represents an issue of undeniable importance, especially relevant for platforms where contributions are made by the end users (Hale & Eleta, 2017). Thus, online content is not only generated and consumed in different languages: in the 2.0 digital environment, new multilingual practices are emerging, and new global communication challenges unfold. Even though it remains to be seen how multilingualism will be specifically integrated in different 2.0 platforms in the coming years, and whether a specific language policy will prevail, the analysis of the current multilingual strategies on 2.0 platforms can indicate possible future developments.

In this paper, I shed some light on the multilingual dimension of the 2.0 Internet. First, in Section 2 I offer a brief overview of the most relevant research conducted in literature connected to the 'multilingual Internet'. Secondly, Section 3 focuses on the multilingual dimension of five of the most popular 2.0 platforms, discussing how these major social networks 'deal' with multilingualism in practical terms, delving into their language policies and incorporated online translation tools, as well as providing a literature review of this emerging research theme. Building on this discussion, I also address the question whether a 2.0 platform, such as TripAdvisor, is heading towards more in-group communication between language-specific groups or, instead, towards a channel for global communication. Finally, indications for future research are discussed in section 4.

Multilingualism Online: Main Contributions

In the past two decades, there has been a growing body of research focusing on different facets of multilingualism on the digital environments. By digital environments, "we refer here to digital media platforms that enable the creation, sharing, and exchange of user-generated content and involve social interaction between participants" (Leppänen, Kytölä & Westinen, 2017, p.120). Indeed, social networking sites, as the ones that will be discussed in this contribution, explicitly rely on the concept of mutual exchange of content. From the

early 2000s onwards, an increasing number of investigations focusing on 'multilingual Internet' has been published. A crucial publication suggesting the increased interest in this research area is represented by the volume *The Multilingual Internet* (edited by Danet & Herrings, 2007), which contains a collection of papers discussing multilingualism online.

In more recent years, this research topic, which has continued to expand, has been accurately described by Barton & Lee (2017, p.142), who pointed out how studies of multilingualism online, so far, tend to fall into two major categories: quantitative studies of linguistic diversity online and, the investigation of patterns of mixed-language practices in a specific form of computer-mediated-communication (CMC) or across different forms of CMC. Thus, a first stream of research consisted in studies investigating the presence of specific languages on the Internet, mainly concentrating on the distribution of English and other languages (e.g., Paolillo, 2007). While a second broad cluster of studies focused on a detailed analysis of the language use and interaction among multilingual web users, adopting various framework, such as discourse analytic, pragmatic, ethnographic or (socio)linguistic perspectives (e.g., Androutsopoulos, 2010; 2011; 2013; 2014; 2015; Kytölä, 2014; Lee, 2015; 2017; Leppänen & Peuronen, 2012). Key issues explored in this specific research area are represented by linguistic phenomena as language choice (e.g. Androutsopoulos, 2007; Hinrichs, 2006; Lee, 2007; 2014; Tagg & Seargeant, 2012) and code-switching practices (e.g. Lee, 2017; Leppänen, 2007; 2012; Sebba, 2012; Siebenhaar, 2006) adopted by multilingual users communicating in the digital sphere. For a more detailed and extensive discussion of the literature of this first-wave research on multilingualism and the Internet, the reader can refer to the recent works of Barton and Lee (2017) and Leppänen et al. (2017).

Reflecting on the current phase of globalization and the constant and fast advancement of digital communication technology, we have seen social media practices become transcultural and multilingual to an extent never experienced before (Kytölä, 2016; Leppänen, 2012; Peuronen, 2011). These developments influence not only the use of language by web-users, but also the language policies adopted by the different global social media platforms. As a matter of fact, to date, a less discussed area in literature concerns the following issues, namely: how do different 2.0 platforms choose to 'manage' multilingualism? Do social networks decide to promote or hinder this feature on their platforms? How? These questions point out some significant emerging themes which are at the moment still relatively under-researched. The discussion of these issues represents the core of this contribution and will be elaborated in the next paragraphs.

Multilingual Features on 2.0 Platforms

At present, multilingualism, which is pervading the vast majority of social media and user-generated content platforms, undoubtedly represents a relevant challenge for platform designers. Indeed, language has become an essential factor to be taken into account when

building or improving a 2.0 platform. Platform designers have to decide how to deal with multilingualism, finding and developing the best ways to adapt their platforms in order to accommodate users of multiple languages.

In general, it is important to notice that 2.0 platforms do not limit their platforms to an English-only rule; instead, they opt for multilingual policies and affordances to deal with users who engage with content in multiple languages (Hale, 2016; Lenihan, 2011). Nonetheless, every platform deals with the issue of multilingualism in its own way, which is characterized by its own mode of communication, discourse and incorporated tools (Lee, 2017).

Wikipedia

The digital and collaborative encyclopedia Wikipedia is available in 299 languages, 11 of which contain over 1,000,000 articles (Wikipedia, 2018). Language is used to organize content by relating articles across languages with interlanguage links (Hale, 2012). On closer inspection, though, it is impossible not to notice the major variation among various language editions. Hecht and Gergle (2010), for instance, found little correspondence of coverage, content or length among same entries written in different languages. Moreover, the majority of articles on Wikipedia exist in one language only (Hale & Eleta, 2017). Thus, the overall consensus on the multilingual character of Wikipedia is that while English clearly has a content advantage, a relevant portion of unique information is available in the different language editions (Bao, Hecht, Carton, Quaderi, Horn & Gergle, 2012). Although Wikipedia "embodies an unprecedented repository of world knowledge diversity in which each language edition contains its own cultural viewpoints on a large number of topics" (Bao et al., 2012, p. 1075), various scholars have described Wikipedia's multilingualism and promotion of linguistic diversity as a 'superficial' one (e.g. Hecht & Gergle, 2010; Lee, 2017), since there is no real interaction or dialogue across languages. In fact, language may even become a kind of barrier, separating content and slowing down the transmission of knowledge and information among users of different linguistic backgrounds (Hale & Eleta, 2017).

Facebook

Facebook also markets itself as global and multilingual. First launched with English-only interfaces (Lee, 2017), Facebook initiated its translation process with Spanish a decade ago and has since been translated into several major languages (Lenihan, 2011; O'Brien, 2011). Indeed, Facebook hopes "to support Facebook in the native language of all our users and people who want to use the site" (Desjardins, 2017, p. 23). Taking a closer look at Facebook's multilingual policy, we discover that the platform has been mostly translated collaboratively "by soliciting its 'crowd', that is the users of the platform" (Desjardins, 2017, p. 23). Facebook has become one of the most successful models of crowdsourced

translated global platform, and thanks to the contributions of its users, it has progressively been able to offer translations of its platform in over 75 languages (Snell-Hornby, 2012).

This crowdsourced translation focuses predominantly on the translation of the platform itself, for instance it has been applied to its interfaces, FAQ sections or guidelines (Desjardins, 2017). In order to further stimulate crowdsourced translations, Facebook even launched an app to sustain this purpose: the Translate Facebook App, underlining in its mission statement how "easy [it is] for translators all over the world to help with the project" (Facebook, 2016). Only recently did Facebook start experimenting with online machine translation. Automatic translation options are now provided through the 'see translation' link, for posts written in 'foreign' languages, as in languages different from that chosen by the user as the default language.

Finally, despite the undeniable achievements obtained by Facebook in making the platform multilingual, mostly thanks to the collaboration of its users, this unprecedented use of 'community'/'crowdsourced' translation has also brought to light some ethical issues (Desjardins, 2017). Indeed, several scholars started wondering if the use of unprofessional, unpaid 'translators' is just, and how this could have an impact on the future of professional translators and of the language service industry in general (e.g. Costales, 2011; Dolmaya, 2011; Fuchs, 2015).

Instagram

Instagram was launched in 2010 and currently ranks as one of the most popular social media world-wide with over 700 million users (Statista 2017). As defined by Lee & Chau (2018, p. 22) "Instagram is an image-/video-sharing [platform] where users snap, post, and share images online instantly". Users can also follow other Instagram accounts and view their photos (Matey, 2018). Indeed, Instagram promotes active sharing and social networking, and while posting on Instagram, users may at the same time choose to share images on other social media, such as Facebook and Twitter, possibly reaching larger audiences (Lee & Chau 2018).

As the platform's main functionality consists of users taking photos and posting them online (Matey, 2018), this social media is clearly characterized by a focus on the visual domain. Nonetheless, language-related aspects are progressively gaining importance in the communicative ecology of the platform. A first relevant change in the communicative practice of Instagram is undoubtedly related to the introduction of hashtags in 2011. Hashtags are keywords chosen and defined by the user, prefixed by the # symbol, and "may consist of just one word (#hope) or a string of words written without spaces (#fightfortherighttobefree)" (Lee & Chau 2018, p. 22). Interestingly, hashtags were spread by users before the designers adopted and standardized the practice as an essential characteristic of the platform (Barton, 2015). At the moment, Instagram allows the option of

complementing each post with hashtags (to a maximum number of 30 hashtags per post) (Lee & Chau, 2018).

A second major change in the communicative practice of this social media platform is related to its growing multilingual dimension. As a matter of fact, the Instagram user base community has not only grown extremely fast, it has also reached a global status, counting subscribers from all over the world among its members. In 2016, the platform supported content written in 25 languages, while as of version 8.4 (2018), the number of supported languages has already increased to 40 (Instagram, 2018). In an effort to make it easier for its users to appreciate content from all over the world, sharing posts and following members using different languages, in 2016 the platform decided to incorporate an online translation tool.

The in-built translation tool translates text in bios (the text used by the user to describe him/herself), but also captions relating to, and comments on, the pictures, while hashtags cannot be translated. In practice, a "see translation" link has been added under the text that can be translated, providing a machine translation in the preferred language of the user (the language chosen in their settings). The automatic translation link initially supported major languages, such as French, German, Italian, Russian and Spanish, but is constantly expanding, progressively including "smaller" languages, such as Danish or Afrikaans (Hypertext, 2016).

In its attempt to accommodate an increasing number of its "global" users, in 2017, Instagram decided to reconfigure the entire platform to support languages such as Arabic, Farsi and Hebrew (Instagram, 2017), which are three of the most used languages written from right to left, therefore more complex to include as screen languages of the platform (TechCrunch, 2017).

Multilingualism on Travel Platforms

Travel and tourism has become one of the largest and fastest-growing economic sectors world-wide (UNWTO, 2015). The growing success of the tourism business is not limited to the 'offline world' but it is mirrored in the online environment. As a matter of fact, tourism and travel are such widely discussed topics in the Web 2.0 setting that the application of the Web 2.0 to the tourism sector has even received its own label and is now often referred to as Travel 2.0 or e-Tourism (Minazzi, 2015).

Nowadays, users are given the opportunity to share their own views, comments and suggestions in an informal and collaborative way. Hypothetically, every product and service could be rated and commented online by its consumers and, as pointed out by Vásquez (2014, p.1), at present "the number of consumer reviews posted on the Internet has exploded and, as a result, today there are literally billions of reviews that can be found on a variety of websites". The exponential growth of consumers' reviews can certainly be witnessed in the field of tourism as well, and in this sector, they are perceived as even more valuable since

they concern intangible "experience" goods (Levy, Duan, & Boo, 2013). More specifically, one of the most consulted, and at the same time produced, forms of user-generated content of the Travel 2.0 environment is "online travel reviews", as they "represent people's wish to share their travel experiences online, recommend a tourism product/service or complain about it" (De Ascaniis & Gretzel 2013, p.157).

At the moment, the two most popular tourism-focused platforms are Booking.com and TripAdvisor, which are examined below. These platforms constitute essential sources of pre-purchase information for travelers and they greatly influence the travel decision-making process of its users (Zeher, Crotts & Magnini, 2011). In addition, they represent the digital repository of travelers' experiences and opinions.

Both platforms have reached such a global status, reaching tourists from all over the world, that they are inevitably confronted with the challenge of multilingualism and the development of an efficient language policy. Indeed, non-English speakers are travelling more and have also become more active online (Hale & Eleta, 2017). This change is also reflected in the user-generated content of these platforms, as the amount of information produced and consumed on these environments is increasingly multilingual. Since both platforms are aware of the great challenge and opportunity that multilingual users bring to their businesses, they are developing different strategies to adapt their services to users from diverse linguistic backgrounds.

Booking.com

Booking.com started as a small Dutch start-up and now provides information in more than 40 languages. In line with the general Booking.com growth, the platform became aware of the need to offer content in different languages and identified machine translation as the most efficient solution to realize this (Levin, Dhanuka & Khalilov, 2017). More specifically, Booking.com is developing an in-house machine translation tool, as customized as possible for the needs of the platform. For many years, machine translation was primarily adopted by the Booking.com translation team in the post-editing stage, in which automatic translation promoted an increase in productivity during the professional translation procedure (Khalilov, 2018). Nonetheless, in this e-commerce environment, the most desirable application of machine translation is direct publishing of machine-translated content (Khalilov, 2018). Booking.com is moving towards this goal, and machine translation is currently being used for translating property descriptions (hotels, apartments, B&Bs, hostels, etc.) from English into any of the other supported languages (Levin et al., 2017). Nonetheless, other sections of Booking.com content are still displayed in the original language without a machine translation. This is the case, for instance, for short evaluations written by the users. As this language policy is still in progress on Booking.com, it is possible to find content written in various languages on a single web page. Although one language may be set for the interface, different languages may well appear in the tourists' evaluations and in the property

description section. Finally, increasing attention is paid to the concept of 'localization', meaning that all translations produced by Booking.com should provide translated content which is the most suitable and appreciated by the different markets, and thus more effective also form a commercial point of view.

TripAdvisor
Launched in 2000 (Lee, Law & Murphy, 2011), it currently welcomes more than 455 million unique monthly visitors, contains 600 million reviews and opinions covering more than 7.5 million accommodations, restaurants, airlines and attractions (TripAdvisor, 2018). Among the specialized customer-review sites in the field of tourism, TripAdvisor is certainly the most popular one. TripAdvisor embodies the largest travel platform where people share knowledge, information and advice about travel services, and it is "one of the first reviews sites to exploit user-generated content" (Vásquez, 2014, p. 8). Starting, as most of the other social networks, as a heavily Anglo-American centered platform, it progressively augmented its supported languages. Indeed, while in 2010, it was possible to read and write reviews in 16 languages and the number of languages increased to 21 by 2013 (Yoo, Sigala & Gretzel, 2016), at present, TripAdvisor provides content in more than 30 languages. This is not a random number but represents the "languages associated with all countries supported by TripAdvisor points of sale" (TripAdvisor, 2018).

The case of TripAdvisor is particularly striking, since it is possible to notice how two opposite strategies dealing with multilingualism are put in place on the same platform (Cenni & Goethals, 2017). Notably, language is a key element in the interactive design of TripAdvisor. On the one hand, users can limit the reading and writing of TripAdvisor reviews to their mother tongue, setting their native language as the preferred one, determining, in this way, a specific review display order. Inevitably, this practice facilitates interactions between users with the same linguistic background (Cenni & Goethals, 2017; Hale, 2016). On the other hand, TripAdvisor incorporated one of the most popular machine-translation tools: Google Translate, which allows users to access and understand reviews written in 'foreign' languages, encouraging a more global approach to the communication patterns among the platform users. This example of dual language policy, detected simultaneously on a single social media platform, gives rise to a compelling question related to multilingualism in the Web 2.0 setting, namely whether 2.0 platforms, and TripAdvisor in particular, could further evolve in the direction of linguistically-separated speech communities or in the direction of a globalized one (Cenni & Goethals, 2017).

The constant improvement of online translation tools and the growing awareness of its added value could possibly suggest a trend in the direction of a more global communication mode that dominates 2.0 platforms, especially tourism-focused ones. As a matter of fact, machine translation represents a tool faster and more affordable than ever, able to provide translation immediately as soon as content appears online (Levin et al., 2017). Additionally,

machine translation significantly improves information retrieval (Hale, 2014; Hale & Eleta, 2017). Indeed, speakers of smaller-sized languages (such as Dutch or Swedish, for instance) are not limited in their search of content by their native language, but thanks to online translation, they have the opportunity to engage with foreign-language content and reach information written in the other supported languages. At the same time, those speakers can still use their native language and know that their contribution can be read by other users because of the machine translation tool (Cenni & Goethals, 2017). In the specific case of tourism-focused platforms, this means that tourists never had better and easier access to travel information and guest reviews. In other words, this development enhances the possibility of making informed choices and purchases before your trip.

Finally, it is important to highlight that, through the adoption of machine translation, users are able to interact with multilingual content and, consequently, mutual intelligibility (Lee, 2017), and global communication modes get highly promoted.

Conclusions and Directions for Future Research

In this contribution, I have presented an overview of the main studies that investigate multilingualism in the digital environment. I paid special attention to the (multi)language policies and incorporated translation tools of five major 2.0 platforms, namely Wikipedia, Facebook, Instagram, Booking.com and TripAdvisor. Combining the results from the main research in this area with observations on the latest developments of the platforms themselves, I aimed to stimulate reflection on how the multilingual dimension of these platforms is evolving. The analysis shows clearly that for all five platforms multilingualism has become one of their essential features; and this in contrast to the strong Anglo-American linguistic bias of an earlier period. However, the journey towards multilingualism has been, and continues to be, different for each platform. For instance, Wikipedia and Facebook mostly relied on the contributions and translations by its users, while Instagram, Facebook and TripAdvisor opted for the introduction of online machine translation tools.

Global platforms provide a communication environment in which content is generated from a wide variety of cultural and linguistic backgrounds. Each of these platforms has therefore, developed different kinds of facilities and constraints for multilingual communication, taking into account both their own mission and their user-base. Undoubtedly, platforms have a vested interest in making and keeping their platforms multilingual. Making a platform multilingual is synonymous with making it accessible and thus, ultimately, profitable. At the same time, meeting the linguistic needs of a greater portion of users proves to be positive, not only for service providers, but also for the users, who can participate in the digital arena, consuming and producing content in a multiplicity of languages.

Reflecting on the platforms' journey towards an enhanced multilingualism, we are able to attest to diverging tendencies (Kelly-Holmes & Lenihan, 2017). On the one hand, we

are witnessing an emergence and growth of top-down multilingualism, as evidenced by the speed with which social networks are providing a greater variety of language options for their users (Kelly-Holmes, 2013). On the other hand, bottom-up multilingualism has also emerged, whereby more and more native speakers of different languages are becoming users of global platforms and are eager to access content provided not only in their own, but also in other, languages. There is also a conjunction of these two practices online, whereby platform users embody both the "source and resource" for multilingualism online (Kelly-Holmes & Lenihan, 2017), as discussed in the case of Facebook "outsourcing" their translation work on behalf of their communities of users (crowdsourcing).

Building on studies conducted to date, there is a growing demand for longitudinal studies that investigate the evolution of different tools and phenomena. For instance, a relevant area for future research would be to investigate the long-term impact of crowdsourcing on the multilingual dimension of the web (Kelly-Holmes & Lenihan, 2017), in particular relating to smaller-scale languages, which, at present, are the languages that have less access to (machine-)translated content. In addition, future studies might investigate how the different multilingual affordances are perceived and made use of by the actual users, thus focusing on the users' perspective.

Finally, platform designers join scholars in realizing how integrating machine-translation tools in 2.0 platforms may dramatically increase translation efficiency and, at the same time, substantially boost intercultural communication. Therefore, it is of fundamental importance to monitor the development of online machine-translation tools adopted by the different platforms, and examine how the evolution of this technology can influence multilingualism on global platforms.

References

Androutsopoulos, J. (2007). Language choice and code- switching in German based diasporic Web forums. In B. Danet & S.C. Herring (Eds.), *The Multilingual Internet: Language, Culture, and Communication Online* (pp. 340-361). Oxford: Oxford University Press.

Androutsopoulos, J. (2010). Localizing the global on the participatory Web. In N. Coupland (Eds.), *The Handbook of Language and Globalization* (pp. 203-231). Oxford: Wiley-Blackwell.

Androutsopoulos, J. (2011). From variation to heteroglossia in the study of computer-mediated discourse. In C. Thurlow & K. Mroczek (Eds.), *Digital Discourse: Language in the new media* (pp. 277-298). Oxford: Oxford University Press.

Androutsopoulos, J. (2013). Participatory culture and metalinguistic discourse: performing and negotiating German dialects on YouTube. In D. Tannen & A. M. Trester (Eds.), *Discourse 2.0. Language and New Media* (pp.47-72). Georgetown University Press.

Androutsopoulos, J. (2014). Moments of sharing: entextualization and linguistic repertoires in social networking. *Journal of Pragmatics 73*, 4-18.

Androutsopoulos, J. (2015). Networked multilingualism: some language practices on Facebook and their implications. *International Journal of Bilingualism 19* (2), 185-205.

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: Bridging the Wikipedia language gap. *CHI '12, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1075–1084), New York.

Barton, D. (2015). Tagging on Flickr as a social practice. In R. Jones, A. Chik, & C. Hafner (Eds.), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age* (pp. 48–65). London: Routledge.

Barton, D. & Lee, C. (2017). Methodologies for researching multilingual online texts and practices. In M. Martin-Jones & D. Martin (Eds.), *Researching Multilingualism: Critical and Ethnographic Perspectives* (pp. 141-154). London: Routledge.

Cenni, I. & Goethals, P. (2017). Negative hotel reviews on TripAdvisor: A cross-linguistic analysis. *Discourse, Context & Media 16*, 22-30.

Costales, A. F. (2011). Crowdsourcing and collaborative translation: Mass phenomena of silent threat to translation studies? *Hermeneus*, 15, 85-110.

Danet, B. & Herring, S.C. (Eds.) (2007). *The Multilingual Internet. Language, Culture, and Communication Online*. Oxford: Oxford University Press.

De Ascaniis, S. & Gretzel, U. (2013). Communicative functions of Online Travel Review titles. A pragmatic and linguistic investigation of destination and attraction TR titles. *Studies in Communication Science 13*, 156-165.

Desjardins, R. (2017). *Translation and Social Media. In Theory, in Training and in Professional Practice*. London: Palgrave.

Dolmaya, J. M (2011). The ethics of crowdsourcing. *Linguistica Antverpiensia*, *10*, 97-111.

Facebook (2016). Translate Facebook App. Retrieved May 6, 2016, from https://www.facebook.com/help/100117036792266

Fuchs, C. (2015). *Culture and economy in the age of social media*. London: Routledge.

Hale, S. A. (2012). Impact of platform design on cross-language information exchange. In Proceedings of the 2012. ACM, CHI EA '12, New York, 1363-1368.

Hale, S.A. (2014). Global Connectivity and Multilinguals in the Twitter Network. In Proceedings of CHI 2014. ACM, New York, 833-842.

Hale, S.A. (2016). User reviews and language: how language influences ratings. In Proceedings of CHI EA 2016, San Jose CA, 1208–1214.

Hale, S.A. & Eleta, I. (2017). Foreign-language Reviews: Help or Hindrance? In: Proceedings of CHI 2017, ACM, 4430-4442.

Hecht, B. & Gergle, D. (2010). The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. In Proceedings of CHI 2010. ACM, New York, 291-300.

Hinrichs, L. (2006). *Codeswitching on the Web: English and Jamaican Creole in Email Communication*. Amsterdam: John Benjamins.

Instagram, 2017. Retrieved May 6 2018, from: https://instagram-press.com/blog/2017/10/31/introducing-instagram-in-arabic-hebrew-and-farsi/

Instagram, 2018. Retrieved May 6 2018, from: https://help.instagram.com/111923612310997?helpref=faq_content

Hypertext, 2016. Instant translation coming to Instagram, but only for Afrikaans here. Retrieved May 4 2018, from: https://www.htxt.co.za/2016/06/23/afrikaans-supported-in-instagrams-captions-and-comment-translations/

Kelly-Holmes, H. (2013). "Choose your language!" Categorisation and control in cyberspace. *Sociolinguistica 27*, 132-145 .

Kelly-Holmes, H. & Lenihan, A. (2017). Virtual ethnographic approaches to researching multilingualism online In: M. Martin-Jones & D. Martin (Eds.). *Researching Multingualism: Criticial and Ethnographic* Perspectives(pp.172-185). Abingdon: Routledge.

Khalilov, M. (2018). Machine translation at Booking.com: what's next?: Proceedings of the *AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing,* 86-143.

Kytölä, S. (2014). Polylingual language use, framing and entextualization in digital discourse: Pseudonyms and 'Signatures' on two Finnish online football forums. In J. Tyrkkö & S. Leppänen (Eds.), *Texts and discourses of new media*. Helsinki: Studies in Variation, Contacts and Change in English.

Kytölä, S. (2016). Translocality. In A. Georgakopoulou & T. Spilioti (Eds.), *Handbook of Language and Digital Communication* (371-388). London: Routledge.

Lee, C. (2007). Linguistic features of email and ICQ instant messaging in Hong Kong. In B. Danet & S.C. Herring (Eds.), *The Multilingual Internet: Language, Culture, and Communication Online* (pp.184-208). Oxford: Oxford University Press.

Lee, C. (2014). Language choice and self- presentation in social media: the case of university students in Hong Kong. In P. Seargeant & C. Tagg (Eds). *The Language of Social Media: Community and Identity on the Internet* (pp. 91-111). London: Palgrave Macmillan.

Lee, C. (2015). Digital discourse@public space: Flows of language online and offline. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), Discourse and digital practices. Doing discourse analysis in the digital age (pp. 175–192). New York: Routledge.

Lee, C. (2017). *Multilingualism Online*. London/New York: Routledge.

Lee, C. & Chau, D. (2018). Language as pride, love, and hate: Archiving emotions through multilingual Instagram hashtags. *Discourse, Context & Media 22,* 21-29.

Lenihan, A. (2011). Join our community of translators: language Ideologies and/in Facebook. In: C. Thurlow & K. Mroczek (Eds.), *Digital Discourse: Language in the new media* (pp. 48-63). Oxford: Oxford University Press.

Leppänen, S. (2007). Youth language in media contexts: insights into the functions of English in Finland. *World Englishes 26* (2), 149-169.

Leppänen, S. (2012). Linguistic and discursive heteroglossia on the translocal internet: the case of web writing. In M. Sebba, S. Mahootian & C. Jonsson (Eds.), *Language Mixing and Code- Switching in Writing: Approaches to Mixed- Language Written Discourse* (233-254). New York: Routledge.

Leppänen, S. & Peuronen, S. (2012). Multilingualism on the internet. In: M. Martin-Jones, A. Blackledge & A. Cresse (Eds.), *The Routledge Handbook of Multilingualism* (pp. 384-402). London: Routledge.

Leppänen, S., S. Kytölä & E. Westinen (2017). Multilingualism and Multimodality in Language Use and Literacies in Digital Environments. In S. L. Thorne & S. May. (Eds.), *Language, Education and Technology* (pp.119-130). Springer International Publishing.

Levin, P., N. Dhanuka & M. Khalilov (2017). Machine Translation at Booking.com: Journey and Lessons Learned. In: Conference Booklet, User Studies and Project/Product Descriptions of the 20th Annual Conference of the European Association for Machine Translation, 29–31 May 2017, Prague, Czech Republic, pp. 81–86.

Levy, S. E., W. Duan, & S. Boo (2013). An Analysis of One-Star Online Reviews and Responses in the Washington, D.C. Lodging Market. *Cornell Hospitality Quarterly 54*(1): 49-63.

Matley, D. (2018). ''This is NOT a #humblebrag, this is just a #brag": the pragmatics of self-praise, hashtags and politeness in Instagram posts. *Discourse, Context Media 22*, 30–38.

Minazzi, R. (2015). *Social Media Marketing in Tourism and Hospitality*. Springer International Publishing Switzerland.

O'Brien, S. (2011). Collaborative translation. In Y. Gambier and L. van Doorslaer (Eds.), *The handbook of translation studies* (pp. 17–20). Amsterdam: John Benjamins.

Paolillo, J.C. (2007). How much multilingualism? Language diversity on the Internet. In B. Danet & S.C. Herring (Eds.), *The Multilingual Internet: Language, Culture, and Communication Online* (408-430). Cambridge: Cambridge University Press.

Peuronen, S. (2011). "Ride hard, live forever": translocal identities in an online community of extreme sports Christians. In C. Thurlow & K. Mroczek (Eds.), *Digital Discourse: Language in the new media* (pp. 154-176). Oxford: Oxford University Press.

Sebba, M. (2012). Researching and theorising multilingual texts. In M. Sebba, S. Mahootian & C. Jonsson (Eds.), *Language Mixing and Code-Switching in Writing* (pp.1-26). New York: Routledge.

Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German internet relay chat rooms. *Journal of Sociolinguistics 10* (4), 481-509.

Snell-Hornby, M. (2012). From the fall of the wall to Facebook: Translation studies in Europe twenty years later. *Perspectives: Studies in Translatology, 20*(3), 365-373.

Statista, 2017. Number of monthly active Instagram users from January 2013 to April 2017. Statista [Website]. Retrieved from: https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users

Tagg, C. & Seargeant, P. (2012). Writing systems at play in Thai-English online interactions. *Writing Systems Research 4* (2), 195-213.

TechCrunch (2017). Retrieved May 28, 2018, from:
https://techcrunch.com/2017/10/31/instagram-now-supports-right-to-left-languages-like-hebrew-and-arabic/

TripAdvisor (2018). Retrieved May 27, 2018, from:
https://developer-tripadvisor.com/content-api/supported-languages/

Vásquez, C. (2014). *The Discourse of Online Consumer Reviews*. London: Bloomsbury.

Wikipedia (2018). Retrieved May 29, 2018, from
https://en.wikipedia.org/wiki/Wikipedia:About#Other_languages

Zehrer, A., J. Crotts & V.P. Magnini (2011). The perceived usefulness of blog postings: an extension of the expectancy-disconfirmation paradigm. *Tourism Management 32*, 106-113.

Validating multilingual hybrid automatic term extraction for search engine optimisation: the use case of EBM-GUIDELINES

Ayla Rigouts Terryn[*] - Ghent University, Belgium
Véronique Hoste - Ghent University, Belgium
Joost Buysschaert - Ghent University, Belgium
Robert Vander Stichele - Ghent University, Belgium
Elise Van Campen - Editor, ebpracticenet, Belgium
Els Lefever - Ghent University, Belgium

ABSTRACT

Tools that automatically extract terms and their equivalents in other languages from parallel corpora can contribute to multilingual professional communication in more than one way. By means of a use case with data from a medical web site with point of care evidence summaries (Ebpracticenet), we illustrate how hybrid multilingual automatic term extraction from parallel corpora works and how it can be used in a practical application such as search engine optimisation. The original aim was to use the result of the extraction to improve the recall of a search engine by allowing automated multilingual searches. Two additional possible applications were found while considering the data: searching via related forms and searching via strongly semantically related words. The second stage of this research was to find the most suitable format for the required manual validation of the raw extraction results and to compare the validation process when performed by a domain expert versus a terminologist.
Keywords: automatic terminology extraction; ATR; terminology.

RESUMEN

Las herramientas que extraen automáticamente términos y sus equivalentes en otros idiomas de corpus paralelos pueden contribuir a la comunicación profesional multilingüe de más de una manera. A través de un caso práctico con datos (extraídos de) ebpracticenet, ilustramos cómo funciona la extracción automática de términos multilingües híbridos a partir de corpus paralelos y cómo se puede utilizar en una aplicación práctica como la optimización de motores de búsqueda. El objetivo original era utilizar el resultado de la extracción para mejorar la recuperación de un motor de búsqueda permitiendo búsquedas multilingües automatizadas. Al considerar los datos, se encontraron dos posibles aplicaciones adicionales: la búsqueda a través de formularios relacionados y la búsqueda a través de palabras muy relacionadas semánticamente. La segunda etapa de esta investigación consistió en encontrar el formato más adecuado para la validación manual necesaria de los resultados de la extracción bruta y comparar el proceso de validación cuando lo realiza un experto en medicina frente a un terminólogo.
Palabras clave: extracción automática de terminología; ATR; terminología.

---

* Corresponding author e-mail: ayla.rigoutsterryn@ugent.be

ACCURATE AND CONSISTENT terminology is essential for professional communication. This has led to the development of terminology management strategies, which often include tools to automate different components of the terminology management workflow. This paper is dedicated to the automatic extraction of multilingual terminology, using a hybrid approach, i.e. a combination of both linguistic and statistical features to identify terminology. The practical use of this strategy will be illustrated by means of a use case for EBM-GUIDELINES, a digital database with 1000 highly structured evidence-based guidelines (point of care evidence summaries), published by DUODECIM, the publishing company of the Finnish General Practitioners. All these guidelines have been translated to English and then to Dutch and French, to enable implementation in Belgium. The aim was to explore the possibilities of automatic term extraction (also known as automatic term recognition or ATR) for the optimisation of search engine recall. Multilingual ATR was performed on parallel corpora in English, French and Dutch. The acquired data inspired three different strategies for search engine optimisation. For each given search term, search engine results can be found containing the search term itself, and, in addition: (1) translations of the search term in different languages, (2) morphological variants of the search term, specifically terms with the same lemma, and (3) terms that are strongly semantically related to the search term. Additionally, auto-completion and auto-suggestion of search terms can be improved with the monolingual lists of automatically extracted terms.

While the ATR method used reached a state-of-the-art performance, the results are not yet perfect and require manual validation before they can be implemented in a search engine. Before moving on to the validation, the data needed to be presented in a suitable format. With regard to terminological validation, there are two commonly used approaches for terminological validation: either the results are validated by a domain expert (without specific training in terminology), or they are validated by a terminologist (without domain expertise). In this case, a domain expert (a medical doctor) was consulted to validate the results of the multilingual term extraction. For this research project, a trained terminologist and translator also validated part of the data for comparison, with identical instructions. Since both validating terms and evaluating translations are known to be highly subjective tasks, it is interesting to consider the impact of the validator's background on this task.

The remainder of this paper is divided into four parts. First, the state-of-the art in the field of monolingual and bilingual term extraction is discussed in section 2. Section 3 describes the term extractor used for these experiments: TExSIS. Section 4 explains how these results might be used for search engine optimisation and includes a short evaluation of the results for that purpose. Section 5 is dedicated to the validation of the results, discussing both the methodology and a comparison of the results by the different annotators. Finally, the results are summarised and interpreted in the conclusion, along with suggestions for further research.

State-of-the-Art

ATR has been a productive field of research within computational linguistics. Early work often focussed on either linguistic (e.g. Bourigault, 1992), or statistical (e.g. Sparck Jones, 1972) clues to search for terms. Linguistically inspired methodologies rely on information such as part-of-speech patterns to identify terms, whereas statistical methods calculate word/term frequencies, often comparing frequencies in a specialised, domain-specific corpus with frequencies in a large, general domain corpus. Kageura and Umino (1996) defined two of the fundamental concepts of automatic terminology extraction: termhood and unithood. Termhood refers to how characteristic or relevant a term is within the researched topic/domain. Unithood describes to which degree multi-word terms form a syntagmatic linguistic unit. Since the linguistic and statistical approaches provide complementary information, later ATR methodologies (Daille, 1994) often combine the two approaches. These are called hybrid methodologies. Another evolution has been the introduction of a multilingual aspect by using parallel corpora to extract equivalents for terms in other languages as well. An example of a hybrid tool for bilingual ATR is TExSIS (Macken, Lefever & Hoste, 2013), which was used for the experiments described in this paper.

The evaluation of ATR has always been rather problematic due to the lack of an unambiguous definition of terms (Rigouts Terryn, Hoste & Lefever, 2018). Terms are generally defined as lexical units which refer to relevant concepts within a specific domain. However, such definitions allow room for interpretation, so human annotators identify terms with a certain measure of subjectivity. Consequently, inter-annotator agreement for term annotation is typically very low.

The two most important measures of ATR accuracy are precision and recall. Precision calculates how many of the automatically extracted candidate terms were evaluated as actual terms by human annotators. Recall measures how many of the terms found by human annotators in a text are also extracted automatically. While precision can be calculated based on the extracted list of terms, the calculation of recall necessitates a fully annotated corpus, large enough to be useful for ATR. Therefore, recall often is not calculated, especially for small-scale research. Both measures can be combined into f-score, which is the harmonic mean of precision and recall. Existing resources such as the IATE (Inter-Active Terminology for Europe) or MeSH (Medical Subject Heading) term banks can be used as a reference. For instance, Laroche and Langlais (2010) use 5000 nominal term pairs from MeSH. However, while using such established resources may decrease subjectivity and annotation effort, they do not reflect recall accurately, since there may always be valid and relevant terms in a corpus that are not present in a term bank. Moreover, since one of the applications of automatic term extraction is to extract new terms to keep these types of term banks up-to-date, this evaluation methodology may miss very relevant terms. Term Evaluator (Inkpen, Paribakht, Faez, & Amjadian, 2016) is a tool designed specifically to evaluate and compare different term extractors. The results of several tools are combined, and the tool provides an

interface in which to efficiently annotate the list of extracted term candidates. While this strategy does not allow the calculation of recall (since only the list of extracted term candidates is annotated, not the terms in the original texts), it does provide the option of calculating *relative recall*, taking the union of all term candidates, extracted by all term extractors, as an approximation of all possible term candidates in the text.

When it comes to the validation of term candidates, generally, a choice needs to be made about whether to have a domain specialist or a terminologist perform the annotation. Involving a domain specialist is not always an option, especially if multiple domains or languages are researched. While inter-annotator agreement scores are sometimes reported, the impact of the annotator on the term validation is rarely researched. Hätty and Schulte im Walde (2018) are a notable exception. They asked 20 laypeople to annotate terms in specialised texts in four different domains: do-it-yourself, cooking, hunting, and chess. The term identification was split into four tasks performed in WebAnno (Yimam, Gurevych, de Castilho & Biemann, 2013): highlighting domain-specific phrases, creating an index, defining unknown words for creating a translation lexicon and creating a glossary. There were seven annotators per task. The authors found that agreement was similar regardless of the task and that "laypeople generally share a common understanding of termhood and term association with domains", but that "laypeople's judgments deteriorate for specific and potentially unknown terms" (Hätty & Schulte im Walde, 2018, p. 325). In another study (Rigouts Terryn, Hoste, & Lefever, accepted), a terminologist annotating terms in different domains reported that, while annotating in a domain for which she was a domain specialist was faster, it can also be more difficult to recognise domain-specific terminology when that terminology has become part of one's general vocabulary.

TExSIS

The ATR tool used for this experiment is TExSIS (Macken, Lefever & Hoste 2013), developed at Ghent University. TExSIS is a hybrid tool which can be used for both monolingual and bilingual term extraction from parallel corpora in English, French, German and Dutch. Given a specialised, domain-specific corpus, TExSIS will first perform a shallow linguistic preprocessing, which includes automatic tokenisation, part-of-speech tagging and lemmatisation. Then, a rule-based linguistic filter extracts all candidate-terms with predefined part-of-speech patterns, both single words and multi-word units. Examples of patterns for English are: noun (e.g. *anaemia*), adjective+noun (e.g. *antiarrythmic* agent), noun+preposition+noun (e.g. *loss of consciousness*) etc. These patterns are, of course, language-dependent. One example of an important difference between languages here is the way compound terms are constructed. In Dutch, compound terms are usually single-word compounds, whereas in French and English, multi-word terms are more common. This directly influences the term extraction, due to the different strategies required for single-word or multi-word term extraction.

The linguistic preprocessing (i.e. identification of candidate terms based on POS-pattern) favours recall over precision, and hence generates too much terms. Therefore, candidate terms are put through a statistical filter. In this phase, several statistical scores are computed to calculate termhood and unithood. Termhood is measured by comparing relative frequencies of candidate terms in the specialised corpus with those in a large, general language corpus, using the term-weighting measure of Vintar (2010). Log-Likelihood Ratio (Rayson & Garside, 2000) is another such termhood measure, which is, in this case, only calculated for single-word terms. C-value (Frantzi & Ananiadou, 1999) was chosen to calculate unithood and for finding nested terms, by looking at the length and the relative frequency of the candidate term itself, versus that of all other candidate terms that enclose this candidate term. The results are ranked based on Vintar's term weighting measure. For the experiment, the cut-off values at this stage were set very low to favour recall.

For multilingual ATR, TExSIS requires a sentence-aligned parallel input corpus. In that case, monolingual ATR will be performed on the two languages separately to generate two monolingual lists of term candidates. To identify equivalent terms in the parallel texts for all candidate terms, automatic word alignment is performed, using GIZA++ (Och & Ney, 2003). Again, the decision was made to favour recall over precision for the translation suggestions.

Besides the termhood (whether the term is relevant to the specialised domain) and unithood (whether separate words belong to a single unit. i.e. a multi-word term) measures, an additional statistic was added for the bilingual component of the ATR: FreqRatio. This metric compares the frequency of the source term candidate and the suggested target term candidate. The intuition behind this metric is, that equivalent terms will probably appear a similar number of times in a parallel corpus. FreqRatio expresses the relative difference in frequency between suggested equivalents and can be used as an additional filter. However, using a hard cut-off based on FreqRatio is not always recommended, since it is very sensitive to differences in frequency caused, e.g. by synonyms and variants.

Multilingual Automatic Term Extraction for EBM-GUIDELINES

Data

EBM-GUIDELINES is a digital database of evidence-based medical guidelines and information for caregivers. Originally in Finnish and English, the database has been translated in Dutch and French, for implementation in Belgium by the company IScientia, using augmented machine translation with a translation memory, and subsequent revision by a professional translator and a medical proof-reader (cf. Van de Velde et al., 2015). The database is accessible online to caregivers through the eHealth Platform and Internet (https://www.ebpnet.be/). An independent non-profit organisation, ebpracticenet, financed by the Belgian government, provides contextualisation of the information for the healthcare system. The texts in this database are written in English, French and Dutch. The EBM-

GUIDELINES are big parallel corpora, providing a large number of aligned translations for English-French and English-Dutch. In addition to the guidelines, the database also contains 5000 English-only summaries of systematic reviews. These reviews underpin recommendations within the guideline. They are based on the work of the Cochrane Collaboration, a worldwide network that specializes in the production and maintenance of systematic reviews of randomized clinical trials in the field of medicine (https://www.cochrane.org/ Last accessed on Dec 20, 2018). This is considered the nec plus ultra of evidence-based medicine.

Dutch-speaking users search the EBM-GUIDELINES using the Dutch interface and Dutch search terms. With the help of an alignment tool coupling search terms in English and Dutch, relevant results can be retrieved, not only from the EBM-GUIDELINES (in Dutch), but also from the English-only Evidence Summaries. Therefore, the term extraction was commissioned by ebpracticenet to improve the search engine recall, both for Dutch and French users. The alignment tool should enable searching across Dutch and English and French and English, so that, for any given search term, the search engine can return both documents containing the search term and documents that contain a translation of the search term.

As input for TExSIS, two sentence-aligned parallel corpora were provided with the translation of nearly one thousand medical guidelines, with an average of 4 pages and 100 aligned segments per document. The source language was English for both corpora, the target languages French and Dutch respectively. Not all English texts were translated in both target languages, so the two parallel corpora are very similar, but not identical in content. The English-French corpus contains 1,101,217 tokens in English and 1,266,731 tokens in French. The English-Dutch corpus contains 1,147,311 English tokens and 1,137,773 tokens in Dutch.

Results from TExSIS

After running TExSIS on both bilingual corpora (English-French and English-Dutch), English was used as a pivot language to create trilingual term lists for the preliminary evaluation. For instance, the English list was based on the English lemmatised candidate terms. Each row contained one lemmatised English candidate term (e.g. *aneurysm*), all full forms of that candidate term found in the corpus (e.g. *aneurysm* and *aneurysms*) and all possible translations of the (lemmatised) candidate term in French and Dutch, including all possible full forms of the suggested translations. Additionally, the information from the term extraction was added: the part-of-speech pattern, frequency and termhood and unithood scores. Separate lists were made based on the French and Dutch lemmatised candidate terms, since not all candidate terms have a version in each language, and some have multiple translations.

Table 1 shows how many different lemmatised candidate terms were found for each language. By presenting all data in sortable tables, the cut-off values could be determined ad-hoc. Since English was used as a pivot language and French and Dutch corpora were not based on exactly the same English corpus, there are more lemmatised candidate terms with one translation in English and all lemmatised candidate terms in French and Dutch have at least one English translation suggestion.

	EN	FR	NL
LEMMATISED CTS WITH MIN. 1 TRANSLATION	74,384	46,408	67,904
LEMMATISED CTS WITH ENGLISH TRANSLATION	n.a.	46,408	67,904
LEMMATISED CTS WITH FRENCH TRANSLATION	45,512	n.a.	40,215
LEMMATISED CTS WITH DUTCH TRANSLATION	64,113	37,012	n.a.

Table 1. Number of extracted lemmatised candidate terms (CTs); n.a. = not applicable.

The data revealed that only a small percentage of all lemmatised candidate terms appear with more than one full form in the corpus: 4-6%. However, since there are so many extracted terms, this still amounts to over ten thousand lemmatised candidate terms with multiple full forms in total. Moreover, these are often important and/or frequent terms, such as *patient*, *symptom* and other common medical occurrences such as *arrhythmia*, *haemorrhage* and *thrombosis*.

To check the relevance of the data for the improvement of the search engine, spot-checks were performed to calculate precision at different points in the ranked list (sorted on Vintar's termhood score). These checks were performed on the English list. A candidate term was considered correct if (1) it was related to the medical domain and (2) could conceivably be used as a search term on the ebpracticenet website. To clarify, we did not evaluate termhood, but potential relevance as a search term in the ebpracticenet search engine. For instance, *insulin requirement of basal metabolism* could be used as a search term but, typically, *insulin requirement* and *basal metabolism* would be considered terms separately.

Evaluation of Results

To compare accuracy in relation to rank (based on termhood measure), 50 terms were annotated at 7 different points: the first 50 terms, then 50 terms at 5%, 10%, 25%, 50% and 75% of the total termhood ranking and the 50 bottom-ranked terms. In total, this resulted in annotations for 350 candidate terms. Inter-annotator agreement was calculated to ensure a nuanced interpretation of the results. The two annotators agreed 85% of the time, resulting in a Cohen's kappa score of 0.6.

Termhood rank	1%	5%	10%	25%	50%	75%	99%	Total
Nr. of analysed terms	50	50	50	50	50	50	50	350
Validated	41	45	42	41	39	22	10	240
DISCARDED	9	5	5	5	5	24	27	80
Named Entity	0	0	3	4	6	4	13	30
PRECISION (INCL. NES)	82%	90%	90%	90%	78%	52%	46%	77%
PRECISION (EXCL. NES)	82%	90%	84%	82%	90%	44%	20%	69%

Table 2. Precision at different termhood ranks (50 terms per percentile)

The results of the evaluation are presented in Table 3. First of all, we see a very high precision for the first half of the candidate terms. Even at the 75th and 99th percentile, up to half of the candidate terms could be relevant, especially when NEs are considered. The first explanation for the quality of these results is that we evaluated usefulness as search terms, not termhood. The evaluation was also lenient by allowing relevant parts of potential search terms: e.g. *failure*, which can be combined in terms such as *organ failure* or *heart failure* but would not be considered a medical term on its own. Despite the limited scope of the evaluation, the results are convincing enough to indicate the practical use of ATR for selecting search terms.

Precision was also calculated for the automatically generated translation suggestions. All previously validated terms were evaluated with respect to the French and Dutch translation suggestions. Named entities and rejected terms were excluded from this analysis. All translation suggestions that were equivalent or nearly equivalent in meaning to the source term were validated. Translation suggestions of a different word class than the source term, but with the same general meaning were also validated (e.g. if the source term was *ill* (adjective), translations of *illness* (noun) were validated as well). Otherwise, the evaluation was very strict, discarding any hyponyms, hypernyms and other strongly related but not synonymous terms. In some cases, a French or Dutch text contained English terminology. These were also discarded, as well as any misspellings. The results of this analysis are presented in Table 4.

	FR	NL
# VALIDATED ENGLISH SEARCH TERMS WITH TRANSLATION(S), OUT OF A TOTAL OF 350 CANDIDATE TERMS	162/350	208/350
% of validated terms with min. 1 correct translation	97%	98%
% of validated terms with only correct translations	81%	82%
% of validated terms with multiple correct translations	22%	24%
Average % of correct translations	89%	88%

Table 3. Precision of translation suggestions

Once again, the results look promising, with nearly all search terms having at least 1 good equivalent in the other languages. There were fewer equivalents in French, since that parallel corpus was smaller, so some of the English terms simply did not occur in the English-French parallel corpus. A large proportion of all search terms have multiple translation suggestions, though not all of the suggested translations are correct. Highly ranked terms are often frequent terms, for which many potential translations are found. For instance, the term disease has 18 different translation suggestions in French and 26 in Dutch. While these lists contain correct translations (e.g. *maladie* in French and *ziekte* in Dutch), they also contain many incorrect suggestions. Translations that were judged as incorrect include the original English form *disease* (instead of a Dutch equivalent), semantically related, but non-equivalent terms such as *problème*/*problem* (EN: *problem*) and *infection/infectie* (EN: *infection*), hyponyms such as the translations *dementia* and *lung infection*. In Dutch, there are also a few complex compound terms, which contain the correct translation, but only as part of the compound, e.g. *ziekteverloop* (EN: *course of the illness*). These were considered incorrect as well. The example of *beta-blockers* reveals another type of related terms: different spellings, e.g. in Dutch: *bètablokker*, *beta-blokker* and □-*blokker*. One more peculiarity we observed was, that, the more general the source term, the more diverse (and inaccurate) the translations. Rarer terms usually have only one, often correct translation suggestion. More general terms, such as *patient* or *disease,* appear very often (creating more room for mistakes) and are regularly translated less literally. For instance, a translator may choose to translate *patient* by *child* if, in a certain context, the two would clearly refer to the same person. In that case, TExSIS may, correctly, identify *child* as the translation, even though they are not equivalents in most cases. Finally, we also noticed how translation suggestions for these terms are often lists of synonyms or alternative spellings, e.g. the translations for *cough medicine*: *antitussive* and *médicament contre la toux* (French) and *hoestmiddel*, *hoestmedicijn* and *hoestmedicatie* (Dutch).

Application: Search Engine Optimisation

There are several ways in which these results, once validated, could contribute to an improved search engine. First and foremost, by allowing multilingual queries, e.g. where a search for *hartfalen* in Dutch would automatically search for *heart failure* in English as well. Second, variants of the same lemma can be searched, so that, e.g. *beta-blockers* would also return results for *beta-blocker*. The third and most difficult application would be to automatically look for strongly semantically related terms, which have the same translation. These data are less accurate, but may be worth considering for very common terms, such as *medicatie, geneesmiddel* and *medicijn* in Dutch. While we have not explored this option in any detail yet, our results do indicate that this may be an interesting next step. Finally, auto-completing terminology based on the known terminology could help users to formulate more relevant queries.

However, before any of these may be implemented, the results need to be manually validated to ensure reliable user-experience. The first step towards this goal is to present the data in a suitable format and to formulate strategies for efficient validation.

Validation by Domain Expert versus Terminologist

Format

The two main requests from ebpracticenet for the validation of the results were: (1) have one term candidate and equivalent suggestion per line and, (2) strategies to quickly eliminate/validate larger batches of term pairs. The former meant that three different bilingual lists needed to be created, each time choosing the *source* language. The three resulting lists are: English – French, English – Dutch, and French – Dutch. The results are based on the full form (not lemmatised) term candidates of the first (source) language. There is only one suggested equivalent in the target language per line, meaning that a single source term candidate may be repeated on several lines, once for each different suggested target language equivalent. The translations are linked based on the lemma, so if multiple full forms exist for a suggested equivalent, there will be multiple rows with one full form each. The latter requirement meant that the part-of-speech patterns, frequencies and all termhood and unithood measures were reported for both the source and target language term candidate, as well as the FreqRatio for the translation pair, as explained in section 3. The following list is an example of one row in the final English – Dutch table:

1. English full form: *beta blockers*
2. Dutch full form: *bètablokker*
3. English lemma: *beta blocker*
4. Dutch lemma: *bètablokker*
5. English POS: *singular noun*
6. Dutch POS: *singular noun*
7. Named Entity tags for the tokens of the English candidate term: *0 0*
8. Named Entity tags for the tokens of the Dutch candidate term: *0*
9. Length of English candidate term (in tokens): *2*
10. Length of Dutch candidate term (in tokens): *1*
11. Frequency of English candidate term: *36*
12. Frequency of Dutch candidate term: *179*
13. Vintar's termhood score for English candidate term: *12.5*
14. Vintar's termhood score for Dutch candidate term: *95.1*
15. C-Value for English candidate term: *35.0*
16. C-Value for Dutch candidate term: *0.25*
17. Log-likelihood ratio for English candidate term: *0*[1]
18. Log-likelihood ratio for Dutch candidate term: *1633*
19. FreqRatio: 397%

As can be seen in this example, the FreqRatio is quite high, even though the translation is correct. This is due to the fact that there are many different forms of this term in both languages and, in English, the term *beta blockers* appears more often with a hyphen: *beta-blockers*, while, in Dutch, the suggested form *bètablokkers* is the most common variant.

To automatically reduce the size of these lists before the manual validation process, a filter was created based on discussions with ebpracticenet about their preferences. Rather than simply filtering on, e.g. Vintar's termhood measure, the term length and frequency were also taken into account. Very long candidate terms with low termhood scores are rarely good terms and very infrequent terms with low termhood scores are rarely relevant. For instance, all terms with a termhood lower than one, were deleted. There were also filters that combined features, e.g. all terms where the product of the termhood score and the frequency was lower than 2.5 were discarded. These filters were determined experimentally and tuned so that, when applied to the English corpus, around 20k unique English lemmas remained. The table was accompanied by explanations about each column, including how they might be used to efficiently validate the results.

Annotation and Results

The actual validation was performed by a domain specialist (medical doctor), who is fluent in all three languages, but has no background in linguistics or terminology. For comparison, a trained translator and terminologist, fluent in all three languages, also performed a part of the validation. Both received the exact same instructions before the task and did not have access to each other's annotations. The instructions by ebpracticenet were not very specific. They wanted correct translation pairs of potentially relevant search terms to be used in their search engine and they wanted only translations in the same full form (e.g. for the English term *aetiology*, the Dutch term *etiologie* could be considered a good equivalent, but not the plural form *etiologieën*). While they hinted at wanting to make a glossary as well, the main purpose was to find relevant and correct translation pairs to improve the search engine. There were no specific instructions on how to deal with items like named entities, so the annotators developed their own strategies according to what they found logical. The annotations were only performed on the English-Dutch data. In total, a sample of 10,000 lines (with one English term candidate and one suggestion for a Dutch equivalent per line) was annotated by both annotators.

The resulting inter-annotator agreement is displayed in Table 5. In 88% of the cases, the annotators agreed, leaving 12% of the lines with different validations per annotator. Both annotators validated over half of the lines and the terminologist validated slightly more than the domain specialist. The resulting Cohen's kappa score for inter-annotator agreement is 0.75. Since evaluation of both terms and translations is notoriously difficult and subjective, this is a relatively high agreement.

	Domain specialist: valid	Domain specialist: not valid	Total
Terminologist: valid	4907	205	**5112**
Terminologist: not valid	1028	3860	**1233**
Total	**5935**	**4065**	**10000**

Table 5. Inter annotator agreement between terminologist and domain specialist

While annotating, the terminologist assigned the data into twelve different categories (see Table 6). Even though these categories are, of course, somewhat subjective, since they are based on the terminologist's assessment, they do allow for a more detailed analysis of the results. They also helped the terminologist to annotate more consistently and make the same decision for similar cases. Another difference in the annotation process between the terminologist and the domain specialist was, that the terminologist sorted the term candidates alphabetically (to easily group the same or similar terms and make a consistent decision) and the domain specialist sorted on termhood score (to prioritise the most relevant terms).

Annotation categories (by terminologist)	Total # terms	Terminologist	Domain specialist: valid	not valid
Correct	5264	valid	4796	468
Incorrect	1606	not valid	9	1597
Same lemma, different full form	1685	not valid	16	1669
Incorrect but strongly related	175	not valid	4	171
Correct but number debatable	128	valid	4	124
Dutch = English	218	not valid	86	132
Not medical or relevant	369	not valid	87	282
Debatable	64	56 valid; 8 not	8	56
Named Entity: brand/medicine	21	valid	5	16
Named Entity: organisation	27	valid	17	10
Named Entity: person with initials	136	valid	51	85
Named Entity: person without initials	307	valid	29	278

Table 6. Categories of annotation as determined by terminologist

Based on the information in Table 6, the annotations were further analysed. Out of 5264 annotations which were considered correct by the terminologist, only 468 (±9%) were not validated by the domain specialist. In many of those cases, not annotating them was probably a simple result of human error. This is supported by the fact that, in at least 50 of these cases, a different full form of the same term pair was annotated as valid, e.g. the domain specialist annotated *bijwerking(en)* as a Dutch equivalent for *adverse event(s)* as valid in the plural form, but not in the singular. Similarly, the mistake may have been made on the side of the terminologist, causing more disagreement. Sometimes, the terminologists

could use her experience with terminology and translation to recognise less logical or more obscure translations, such as *knutten* as a translation for *biting midges*, *Alzheimer* as a correct translation for *Alzheimer's disease*, even without the explicit addition of a translation for *disease*, or recognising *spm* as a valid translation for *bpm* (abbreviation for *beats per minute)*. In other instances, the terminologist lacked the necessary domain expertise to easily recognise specialised terms, e.g. recognising that the Dutch term *sartanen* is a synonym for *angiotensine receptorblokkers* or knowing whether *arterial disease* can be translated as *vaatziekte* (literally *vascular disease*) or if these are different diseases.

In only nine cases did the domain specialist not agree on a term pair deemed incorrect by the terminologist, including small mistakes made by both annotators. More interesting categories are the next three: *same lemma different full form, incorrect but strongly related* (both semantically and morphologically, e.g. same concept but different part of speech) and *correct but number debatable* (e.g. when a singular term expresses the same meaning as a plural term in the other language). The task was to only consider term pairs correct when they are in the same form. Overall, the terminologist seems to have had the advantage in this case, annotating more consistently. Some of these differences are also due to how strict the instructions were interpreted. For instance, a term like *medication* (or *medicatie* in Dutch) is singular, so is it a correct translation for *drug* (*geneesmiddel)* but not for *drugs* (*geneesmiddelen*)? The domain specialist only validated the two singular forms, whereas the terminologist validated both. She reasoned that *medication* could be used to describe a collection of more than one *drug* and that, in the case of translations, *medication* could often be used for both the singular and plural forms. Similarly, for terms which can be used in plural but rarely are, e.g. *pain(s), discomfort(s), bleeding(s), tendency/-ies,* etc. The terminologist's strategy in these cases was to err on the side of leniency, while the domain specialist tended to only approve term pairs with the same grammatical number. This is also reflected in the category *debatable*, which contains term pairs that are perhaps not literal translations but could, in many cases, be used as equivalents. An example would be *oorzaak* (literally: *cause)* as an equivalent for *aetiological factor*. These two are not synonyms but, in the context of medical texts, can sometimes be used for the same concept. Another example is *coagulation*, which, technically, can refer to other concepts than *blood coagulation*, but in the context of these medical texts, it may be fair to assume they can be used as synonyms.

An exception to this pattern of leniency for the terminologist versus strictness of the domain specialist is in the case of untranslated (English) terms in the Dutch text. The terminologist only approved untranslated terms when they were Named Entities or so common in Dutch that they are used more or equally regularly than the actual Dutch term. For instance, the suggested Dutch translations for *ACE-inhibitor* were *ACE-remmer* or *ACE-inhibitor*. The latter variant may appear to be an untranslated English term at first sight, yet it is common enough to also appear with a Dutch plural form: *ACE-inhibitoren*, rather than the English plural *ACE-inhibitors*. This led the terminologist to accept both *ACE-remmer* and

ACE-inhibitor as valid Dutch translations. The domain specialist also rejects some very clearly untranslated terms but is less consistent when the difference is more subtle, e.g. an ending in -y instead of -ie, or a plural in -s instead of -en. The final categories to discuss are the Named Entities, for which we distinguish between organisations, brands and personal names. The terminologist decided to approve all of the above for the sake of consistency. The translations may not be very informative, since source and target terms should be identical, but the named entities can still be relevant search terms. The domain specialist generally rejected named entities (especially person names), but with many exceptions.

Overall, it appears that the annotators tackled this task with slightly different mindsets. The terminologist was less strict and more likely to approve non-literal translations than the domain specialist, who was stricter, except for non-translated (English) terms as Dutch equivalents. In conclusion, both the terminologist and the domain specialists had advantages and disadvantages. While the domain specialist was able to identify correct term pairs for specialized medical concepts more efficiently, the terminologist could use her experience to annotate more consistently and make informed decisions about term pairs which are less obviously equivalent. Some of these differences are, of course, due to the very general instructions for this task. Ideally, decisions such as whether to annotate proper names should be made beforehand by the client. Moreover, since only two annotators participated in this comparison, we should be careful about generalising these results. Still, the results suggest that the two annotation styles are complementary.

Conclusion

In this paper, we have shown how multilingual automatic term extraction from a parallel corpus has potential for a real-world application such as search engine optimisation. We showed how, despite the need for manual validation, ATE can efficiently produce a list of candidate terms that contains many relevant search terms and, for most of these, good equivalents are found automatically in the other languages. Four potential applications were suggested: (1) multilingual searches, (2) autocompletion of search terms, (3) searching for morphologically related forms using the automatic lemmatisation, and (4) searching for semantically related forms by clustering multiple translations (and back-translations) for the same candidate term.

A suitable format for validation was developed based on feedback from ebpracticenet, who are currently validating the dataset for the implementation of the first application. For this validation, they chose to collaborate with a domain specialist (a medical doctor), who is fluent in the three languages. This is a common strategy, since it is often assumed that domain expertise is necessary to efficiently manage terminology. However, we asked an experienced terminologist without domain expertise to also validate a sample of the dataset. This resulted in 10,000 shared annotations to compare. It was found that, while domain expertise can be an advantage in the case of very specialised terms, the experienced

terminologist was able to annotate more consistently. The domain specialist was generally stricter with the validation, but, since there were only two participants, it is unclear whether this was caused by the lack of detail in the instructions, which left room for the usual subjectivity of this task, or, whether these differences were due to the different backgrounds of the annotators. This would be an interesting path to investigate further, so that users may make a better informed and motivated decision about the person best suited for their validation task.

Even this small-scale study already suggests that, ideally, validation of the results of multilingual automatic term extraction for a real-world application such as search engine optimisation would happen in a multidisciplinary setting, i.e. involving both a terminologist and a domain specialist. Clear instructions should be determined beforehand, preferably combining the input of the client and both a terminologist and domain specialist. One strategy would be to start by having both annotators validate a small subsample and analysing the results and differences to formulate the most suitable strategy for the remainder of the task. Whichever strategy is preferred, the validation will likely benefit from the complementary skills of both a domain specialist and a terminologist.

Notes

1. Since log-likelihood ratio is only calculated for single-word terms, the English multi-word term receives a score of zero.

References

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. *Proceedings of the 14th Conference on Computational Linguistics-Volume 3*, 977–981. Association for Computational Linguistics.

Carroll, J. M., & Roeloffs, R. (1969). Computer selection of keywords using word-frequency analysis. *Journal of the Association for Information Science and Technology*, *20*(3), 227–233.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J. Klavans & P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Aproaches to Language* (pp. 49–66). Massachusetts: MIT Press.

Frantzi, K. T., & Ananiadou, S. (1999). The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, *6*(3), 145–179.

Hätty, A., & Schulte im Walde, S. (2018). A Laypeople Study on Terminology Identification across Domains and Task Definitions. *Proceedings of NAACL-HLT 2018*, 321–326. New Orleans, USA: ACL.

Inkpen, D., Paribakht, T. S., Faez, F., & Amjadian, E. (2016). Term Evaluator: A Tool for Terminology Annotation and Evaluation. *International Journal of Computational Linguistics and Applications*, *7*(2), 145–165.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition. *Terminology*, *3*(2), 259–289.

Laroche, A., & Langlais, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 617–625. Beijing, China.

Macken, L., Lefever, E., & Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, *19*(1), 1–30.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19–51.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, 38th Annual Meeting of the Association for Computational Linguistics*, 1–6. Hong Kong, China: Association for Computational Linguistics.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2018). A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. *Proceedings of LREC 2018*. Presented at the Miyazaki, Japan. Miyazaki, Japan: ELRA.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2019). In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, 1–34. https://doi.org/10.1007/s10579-019-09453-9

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Van de Velde, S., Macken, L., Vanneste, K., Goossens, M., Vanschoenbeek, J., Aertgeerts, B., Vanopstal, K., et al. (2015). Technology for large-scale translation of clinical practice guidelines : a pilot study of the performance of a hybrid human and computer-assisted approach. *JMIR Medical Informatics*, 3(4).

Vintar, S. (2010). Bilingual Term Recognition Revisited. *Terminology*, *16*(2), 141–158.

Yimam, S. M., Gurevych, I., de Castilho, R. E., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6. Sofia, Bulgaria: ACL.

High-density knowledge rich contexts

Pilar León-Araúz[*] - University of Granada, Spain
Arianne Reimerink - University of Granada, Spain

ABSTRACT

Knowledge Rich Contexts (KRCs) are one of the usual data categories contained in terminological knowledge bases. In this paper we show how to extract KRCs that combine various items of knowledge to facilitate the work of terminographers as well as user knowledge acquisition. For this reason, a new knowledge pattern-based sketch grammar was designed within the corpus analysis tool Sketch Engine and applied to the EcoLexicon Environmental Corpus. After compiling the corpus with the sketch grammar, KRCs can be extracted through customized semantic word sketches (WS), which provide access to concordances where the item queried is related to others through one or several semantic relations. Then new queries are performed, reusing these new WS in order to collect high-density KRCs, which combine different semantic relations in a single sentence. In this paper we provide a characterization of high-density KRCs based on the amount of knowledge, concept types, conceptual depth, and number and type of conceptual relations codified in them.
Keywords: knowledge rich context; knowledge pattern; sketch grammar; EcoLexicon Environmental Corpus.

RESUMEN

Los Contextos Ricos en Conocimiento (CRC) son un campo habitual en las bases de conocimiento terminológicas. Los CRC que combinan varios elementos de conocimiento facilitan la labor de los terminógrafos y la adquisición del conocimiento de los usuarios. Por este motivo, se diseñó una nueva gramática colocacional *sketch grammar* para la herramienta de análisis de corpus Sketch Engine que, además de basarse en relaciones gramaticales como los *sketch grammar* por defecto, se basa en patrones de conocimiento y se aplicó al corpus medioambiental EcoLexicon English Corpus. Después de compilar el corpus con la gramática, se pueden extraer CRC utilizando los *word sketches* (WS) personalizados que relacionan el elemento de búsqueda con otros a través de una o varias relaciones semánticas. Después se reutilizan estos WS para realizar nuevas búsquedas y extraer CRC de alta densidad, aquellos que combinan varias relaciones semánticas diferentes en una sola oración. En este artículo se presenta una caracterización de los CRC de alta densidad basada en: la cantidad de conocimiento, los tipos de conceptos, la profundidad conceptual y el número y tipo de relaciones codificados en ellos.
Palabras clave: contexto rico en conocimiento; patrón de conocimiento; *sketch grammar*; EcoLexicon Environmental Corpus

---

* Corresponding author, email: pleon@ugr.es

KNOWLEDGE RICH CONTEXTS (KRCs) (Meyer, 2001) are one of the usual data categories contained in terminological knowledge bases (TKBs), such as EcoLexicon, a multilingual and multimodal TKB on the environment (ecolexicon.ugr.es; Faber et al., 2014; Faber et al., 2016, San Martín et al., 2017). KRCs are conceptually valuable contexts because they contain a term of interest in a particular domain that is semantically related to other terms. KRC extraction is thus essential for terminographic research. One of the most common approaches to find such contexts is to search for the terms in each entry in combination with knowledge patterns (KPs) in corpora. KPs are the linguistic and para-linguistic patterns that convey a specific semantic relation in real texts. For instance, examples of generic-specific KPs are *such as*, *is a kind of, and other*, etc.

KPs are considered one of the most reliable methods for the extraction of semantic relations (Barrière, 2004; Bowker, 2003; Condamines, 2002; Marshman, 2002; to cite a few). They have been applied in many terminology-related projects that have led to the development of knowledge extraction tools, such as Caméléon (Aussenac-Gilles and Jacques, 2008) and TerminoWeb (Barrière and Agbago, 2006).

However, no user-friendly application allowing terminologists to find KRCs in their own corpora is publicly available. The corpus query system, Sketch Engine (Kilgarriff et al., 2004), provides the Word Sketch (WS) function that rather than looking at an arbitrary window of text around the headword – as occurs in previous corpus tools –is able to look for each grammatical relation that the word participates in (Kilgarriff et al., 2004). WS are automatically extracted based on a sketch grammar previously defined with which corpora are compiled. These grammars, based on regular expressions and POS tags, identify and annotate in the corpus different structures of interest. For example, the following grammar rule enables the system to show, in the form of WS, statistically significant structures where a noun is modified by other nouns, adjectives and/or adverbs: 2:"(JJ|N).*" [tag="JJ.?"| tag="RB.?"|word=","]{0,3} "N.*"{0,2} 1:"N.*" [tag!="N.*"]. The default word sketches provided by Sketch Engine represent different relations, such as verb-object, modifiers or prepositional phrases. However, they do not represent conceptual relations. For this reason, in León-Araúz et al. (2016), a KP-based sketch grammar for Sketch Engine was developed, thanks to which a list of KRCs can be obtained when querying any item in a corpus through the word sketch functionality.

For contexts to be regarded as KRCs, they should indicate at least one item of domain knowledge that could be useful for conceptual analysis, whether it is an attribute or relation (Meyer, 2001, p. 281). However, obtaining as many items of domain knowledge in as few KRCs as possible would greatly facilitate both the work of terminographers and user knowledge acquisition. For instance, the following fragments (see Table 1) are high-density KRCs because they combine different KPs (in italics) conveying different semantic relations (i.e. hyponymy, meronymy and causality) between many different concepts (in bold):

> **Nuées ardentes** *are but one type of* **pyroclastic flow**, *which include a variety of mixtures of* **volcanic blocks**, **ash**, **gas**, *and* **lapilli** *that produce* **volcanic rocks** *called* **ignimbrites**.
>
> **Contact metamorphism** *of* **carbonate rocks** *produces* **skarn deposits** *containing* **minerals** *such as* **wollastonite**, **tremolite** and **grossular garnet**, **spessartine garnet** and **andradite garnet**.

Table 1. High-density KRCs

The aim of this paper is to show how, thanks to a KP-based sketch grammar, such KRCs can be collected and analyzed. The focus of this paper is therefore on the characterization of high-density KRCs. Measuring the precision and recall of our sketch grammar is out of the scope of this paper.

The remainder of this paper is organized as follows: "Methodology: Improved KRC Extraction" explains our methodology for improved KRC extraction with our customized sketch grammar; in "Characterizing High-Density KRCs" we apply further queries by reusing the annotated semantic WS to collect different types of high-density KRC and we provide a detailed description of the characteristics of the KRCs collected; finally, some conclusions are drawn and ideas for future work are described in "Conclusions and Future Work".

Methodology: Improved KRC Extraction

The methodology for improved KRC extraction is based on the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz and San Martín 2018; León-Araúz et al., 2016) developed within Sketch Engine and applied to the EcoLexicon English Corpus (EEC) (León-Araúz et al., 2018)[1].

The EEC is a 23.1-million-word corpus of contemporary environmental texts. Each text in the EEC is tagged with a set of XML-based metadata, some of which are based on the Dublin Core Schema[2], while others have been included to meet our own needs. Corpus metadata permit users to constrain corpus queries based on pragmatic factors, such as environmental domains and target reader. Thus, for instance, the use of the same term in different contexts can be compared. Tags are based on the following main parameters:

- Domain: the EEC encompasses all the domains and subdomains of environmental studies (e.g., Biology, Meteorology, Ecology, Environmental Engineering, Environmental Law, etc.).
- User: the corpus includes texts for three types of user, depending on level of expertise (i.e. expert, semi-expert, general public), based on parameters such as the sender (expert in the field, journalist, governmental institution, etc.), type of document (specialized paper, manual, brochure, etc.), place of publication (specialized journal, general website, etc.) and specific mention of intended audience.

- Geographical variant: it comprises American, British, and Euro English (official documents from the European Union).
- Genre: it covers a wide variety of text genres (e.g. journal articles, books, websites, lexicographical material, etc.).
- Editor: it distinguishes texts edited by scholars/researchers, businesses, government bodies, etc.
- Year: it includes texts from 1973 to 2016.

The EEC was compiled with the Penn Treebank tagset (TreeTagger version 3.3) and with the ESSG, a CQL-based (Corpus Query Language) customized sketch grammar also containing the default sketch grammar. So far, the ESSG contains 64 sketch grammar rules[3], focused on the extraction of conceptual relations, which expands the functionality of word sketches to summarized representations of semantic behavior, namely, KRCs. This new sketch grammar for the English language includes some of the most common conceptual relations used in Terminology: generic-specific, part-whole, location, cause, and function.

In the development of the ESSG, issues specific to each relation and pattern had to be taken into account. For instance, a single sentence can produce more than one term pair because of the enumerations that are often found on each side of the pattern (e.g. *x, y, z and other types of w*). This entails performing greedy queries in order to allow any of the enumerated elements fill the target term. However, this may also cause endless noisy loops. Sometimes it is necessary to limit the number of possible words on each side of the pattern. For example, enumerations are more often found on the side of hyponyms, parts, and effects than on the side of hypernyms, wholes, and causes. Consequently, the loops were constrained accordingly in the latter case. Table 2 shows a summarized and simplified version of the patterns included for each semantic relation in the ESSG.

Generic-specific (18 sketch grammars): HYPONYM ,|(|:|is|belongs (to) (a|the|…) type|category|… of HYPERNYM // types|kinds|… of HYPERNYM include|are HYPONYM // types|kinds|… of HYPERNYM range from (…) (to) HYPONYM // HYPERNYM (type|category|…) (,|() ranging (…) (to) HYPONYM // HYPERNYM types|categories|… include HYPONYM // HYPERNYM such as HYPONYM // HYPERNYM including HYPONYM // HYPERNYM ,|(especially|primarily|… HYPONYM // HYPONYM and|or other (types|kinds|…) of HYPERNYM // HYPONYM is defined|classified|… as (a|the|…) (type|kind|…) (of) HYPERNYM // classify|categorize|… (this type|kind|… of) HYPONYM as HYPERNYM // HYPERNYM is classified|categorized in|into (a|the|…) (type|kind|…) (of) HYPONYM // HYPERNYM (,|() (is) divided in|into (…) types|kinds|… :|of HYPONYM // type|kind|… of HYPERNYM (is|,|() known|referred|… (to) (as) HYPONYM // HYPONYM is a HYPERNYM that| which|… // define HYPONYM as (a|the|…) (type|category|…) (of) HYPERNYM // HYPONYM refers to (a|the| …) (type|category|…) (of) HYPERNYM // (a|the|one|two…) (type|category|…) (of) HYPERNYM: HYPONYM

Part-whole (17 sketch grammars): WHOLE is comprised|composed|constituted (in part) of|by PART // WHOLE comprises PART // PART composes WHOLE // PART is|constitutes (a|the|…) part|component|… of WHOLE // WHOLE has|includes|possesses (…) part|component|… (,|() (:|such as|usually|namely|…) PART // WHOLE has|includes|possesses (a|the|…) fraction|amount|percent… of PART // WHOLE part|component|… (,|() such as PART // part|component|… of WHOLE (,|() (:|such as|usually|namely|…) PART // (a|the|one|two| some|…) part|component|… of WHOLE is PART // (a|the|one|two|some|…) part|component|… of WHOLE (is) called|referred|… (to) (as) PART // PART (,|() (a|the|…) part|component|… of WHOLE // WHOLE is divided in| into (two|some|…) parts|components|… (,|() (:|such as|usually|namely|…) PART // WHOLE is divided in|into PART // WHOLE (is,|() made|built|… (up) of|from|with PART // WHOLE contains PART // PART (is) contained in WHOLE // WHOLE consists of PART
Cause (10 sketch grammars): CAUSE (is) responsible for EFFECT // CAUSE causes|produces|… EFFECT // CAUSE leads|contributes|gives (rise) to EFFECT // CAUSE-driven|-induced|-caused EFFECT // EFFECT (is) caused|produced|… by|because|due (of|to) CAUSE // EFFECT derives|results from CAUSE // cause of EFFECT is CAUSE // CAUSE (is) (a|the|…) cause of EFFECT // CAUSE (,|() (a|the|…) cause of EFFECT // EFFECT is|,|(forms|formed by|from CAUSE
Location (4 sketch grammars): ENTITY (is) connected|delimited to|by PLACE // ENTITY (is) found|built|… in|on|… PLACE // ENTITY (is) formed|forms in|on|… PLACE // ENTITY (is) extended|extends (out) into| parallel|… (of|to) PLACE
Function (7 sketch grammars): ENTITY (has|provides|…) (a|the|…) function|role|purpose of FUNCTION // ENTITY is (built|designed|…) for|to FUNCTION // ENTITY is (useful|effective|…) for|to FUNCTION // ENTITY is (a|the|…) (…) built|designed|… for|to FUNCTION // ENTITY is (a|the|…) (…) used|employed|… for|as FUNCTION // use|employ|… ENTITY for|as|to FUNCTION // function|role|purpose of ENTITY is FUNCTION

Table 2. Simplified version of the patterns included in each grammar

Each grammar rule is the formalization of KPs in the form of regular expressions combined with POS tags. As an example, Table 3 shows the actual CQL representation of one of the generic-specific KP-based rules, followed by an explanation and three natural language examples of concordances matched with the grammar.

1:"N.*" [word=",|\\("]? [tag="IN/that|WDT"]? "MD"* [lemma="be|,|\\("] "RB.*"* [word="classified| categori.ed"] ([word="by"] [tag!="V.*"]+)? [word="in|into"] [tag!="V.*"]* [lemma="type|kind| example| group|class| sort|category|family|species|subtype| subfamily|subgroup| subclass|subcategory|subspecies"]? [tag!="V.*"]* 2:[tag="N.*" & lemma!="type|kind|example| group|class| sort|category|family|species| subtype|subfamily|subgroup| subclass|subcategory|subspecies"]	
1:"N.*"	The hypernym is a noun.
[word=",|\\("]?	An optional comma or bracket.
[tag="IN/that|WDT"]?	Optionally "that" or "which".
"MD"*	Any modal verb from zero to infinite times.
[lemma="be|,|\\("]	Lemma "be" or a comma or a bracket.
"RB.*"*	Any adverb from zero to infinite times.
[word="classified|categori.ed"]	Classified, categorised, or categorized.
([word="by"] [tag!="V.*"]+)?	Optionally, "by" followed by anything from one to infinite times that does not contain a verb.
[word="in|into"]	In or into.
[tag!="V.*"]*	Anything from zero to infinite times that does not contain a

	verb.
[lemma="type|kind| example|group|class| sort|category|family|species|subtype| subfamily|subgroup|subclass|subcategory| subspecies"]?	Optionally any of the lemmas "type", "kind", "example", "group", "class", "sort", "family", etc.
[tag!="V.*"]*	Anything from zero to infinite times that does not contain a verb.
2:[tag="N.*" & lemma!="type|kind| example| group|class|sort|category|family| species|subtype|subfamily| subgroup| subclass|subcategory|subspecies"]	The hyponym is any noun other than "type", "kind", "example", "group", "class", "sort", "family", etc.
Stony-iron <u>meteorites</u> are classified into <u>pallasites</u> and <u>mesosiderites.</u> Modern <u>reefs</u> are classified into several geomorphic types: <u>atoll</u>, <u>barrier</u>, <u>fringing</u>, and <u>patch</u>. Littoral <u>materials</u> are classified by grain size in <u>clay</u>, <u>silt</u>, <u>sand</u>, <u>gravel</u>, <u>cobble</u>, and <u>boulder</u>.	

Table 3. CQL representation of a generic-specific KP-based rule with its explanation

Once all the rules in the ESSG are applied, different semantic word sketches can be automatically derived, as shown in Figure 1 for the term *mineral* (only hyponymic and meronymic sketches are displayed).

mineral *(noun)* Alternative PoS: adjective (freq: 127)
EcoLexicon English (Environment) freq = 5,252 (183.53 per million)

"mineral" is the generic of...			"mineral" is part of...			"mineral" is a type of...		
		21.42			10.36			6.80
quartz	38	9.96	rock	79	10.65	mineral	23	8.98
gold	26	9.40	soil	15	8.71	rock	12	8.03
mica	24	9.35	magma	7	8.63	resource	10	7.99
feldspar	24	9.35	mineral	9	8.54	substance	7	7.85
carbonate	22	9.12	melt	6	8.46	earth	7	7.63
iron	24	9.06	jade	6	8.46	material	14	7.51
calcite	19	9.04	peridotite	5	8.18	compound	7	7.47
mineral	23	8.98	silt	5	8.13	soil	6	7.20
copper	17	8.74	crust	6	8.08	feature	8	6.93
clay	17	8.62	meteorite	6	8.07	chemical	5	6.83
salt	16	8.54	limestone	5	7.97	factor	6	5.75
amphibole	13	8.46	planet	5	7.97			
sulfide	13	8.41	type	7	7.91	"mineral" has part...		
calcium	13	8.35	earth	7	7.83			4.70
pyroxene	11	8.28	deposit	5	7.70	silicon	11	10.14
sulfur	12	8.19	material	6	7.60	co3	7	9.68
zinc	11	8.19	sand	5	7.56	carbonate	11	9.61
uranium	11	8.17	sediment	6	7.48	oxygen	14	9.61
magnetite	10	8.14	beach	5	7.09	calcium	8	9.27
coal	13	8.11	water	7	7.07	anion	5	9.25
oxide	11	7.91	group	5	7.02	magnesium	6	9.23
garnet	8	7.83				iron	8	9.17
silver	8	7.78				aluminum	5	9.08
ore	8	7.77				silica	5	8.94
amount	9	7.72				ion	6	8.93
						mineral	9	8.54

Figure 1. Semantic word sketches of *mineral*

When clicking on the number next to each item, concordances unfold as depicted in Figure 2, where *quartz* is shown as a type of *mineral* through different KPs (*such as*, *is a*, *and/or other*, *typically*, *including*).

in a regular three-dimensional pattern to form minerals such as *quartz* (silicon dioxide) or calcite (

' planes of weakness. *Quartz* is a common mineral that lacks cleavage , however. Hardness is

component too , such as *quartz* , pyrite and other minerals , and evaporite minerals , especially

17.18. Continental crust is composed mostly of minerals such as feldspar and *quartz* , which form

reportedly consists primarily of brittle minerals such as calcite , dolomite , albite feldspar ,

being replaced by calcite , *quartz* , or other minerals during burial and diagenesis. Fossils are

conditions fibrous minerals or other minerals , typically *quartz* or calcite , may fill the

depending on the strain rate. Most other common minerals , including *quartz* and calcite , stop behaving

Figure 2. Concordances of QUARTZ *type_of* MINERAL

In Sketch Engine, these concordances can also be shown in the form of whole sentences, thus complying with the idea of KRCs (Figure 3).

Most of the earth 's materials have their molecules arranged in a regular three-dimensional pattern to form minerals such as *quartz* (silicon dioxide) or calcite (calcium carbonate).

Quartz is a common mineral that lacks cleavage , however.

There can be a noncarbonate component too , such as *quartz* , pyrite and other minerals , and evaporite minerals , especially gypsum-anhydrite , can be found in association with limestones.

Continental crust is composed mostly of minerals such as feldspar and *quartz* , which form less-dense , lighter-colored granitic rocks.

The shale matrix reportedly consists primarily of brittle minerals such as calcite , dolomite , albite feldspar , ankerite , *quartz* as well as significant rutile and pyrite.

To be preserved life-forms become mineralized after they die , with organic tissues typically being replaced by calcite , *quartz* , or other minerals during burial and diagenesis.

Under some deformation conditions fibrous minerals or other minerals , typically *quartz* or calcite , may fill the spaces between the boudins.

Most other common minerals , including *quartz* and calcite , stop behaving brittlely and deform ductily at 570oF-1,470oF (300oCâ-800oC).

Figure 3. KRCs for QUARTZ *type_of* MINERAL

As previously mentioned, these can be regarded as KRCs because they indicate, at least, one item of knowledge. However, some KRCs are more useful than others. For instance, the second example in Figure 5 only contains one conceptual proposition[4] (QUARTZ-*type_of*-MINERAL), whereas the fifth activates eight (CALCITE-*type_of*-MINERAL; DOLOMITE-*type_of*-MINERAL; ALBITE-*type_of*-MINERAL; FELDSPAR-*type_of*-MINERAL; ANKERITE-*type_of*-MINERAL; QUARTZ-*type_of*-MINERAL; RUTILE-*type_of*-MINERAL; PYRITE-*type_of*-MINERAL). In these two examples, *type_of* is the only conceptual relation that can be extracted. However, the fourth KRC conveys two: *made_of* (CONTINENTAL CRUST-*made_of*-MINERAL; GRANITIC ROCK-*made_of*-MINERAL) and *type_of* (FELDSPAR-*type_of*-MINERAL; QUARTZ-*type_of*-MINERAL). Therefore, we can assume that, among KRCs, there are poorer KRCs and high-density KRCs. High-density KRCs can be defined as those that contain two or more conceptual propositions.

A high-density KRC can be useful both for users when querying a TKB and for terminographers when building a TKB. Users may activate a whole conceptual network in

their brains if in a single sentence they are confronted with multiple concepts and relations. In the same way, terminographers will populate their TKB in a much more efficient way if a single sentence provides them with multiple conceptual propositions. Moreover, another advantage of high-density KRCs is their reusability, since they can be used in as many term entries as conceptually related terms are found in the KRC.

Characterizing High-Density KRCs

KRCs have been previously characterized in a study conducted by Condamines et al. (2013), where the authors differentiate between conceptually rich and linguistically rich contexts based on their usefulness for translators. In this paper, the focus is on further characterizing high-density KRCs, which can be regarded as a kind of conceptually rich contexts. Three different sets of this kind of KRCs were collected through different queries reusing the semantic WS annotated through the ESSG: (1) KRCs where three different conceptual relations could be found; (2) KRCs codifying two conceptual relation types but several conceptual propositions; and (3) KRCs codifying only one conceptual relation type but several conceptual propositions.

For extracting them, the following query (Figure 4) was first applied to the corpus after excluding the documents of the corpus annotated with Genre "Lexicographic material". This way we avoided confusing definitions from dictionaries and thesauri with actual KRCs in text.

```
([ws(".*-n","\"%w\" is a type of...",".*-n")] []* [ws(".*-n","\"%w\" has part...",".*-n")] []*[ws(".*-
n","\"%w\" is the cause of...",".*-n")] | [ws(".*-n","\"%w\" is a type of...",".*-n")] []*[ws(".*-n","\"%w\"
is the cause of...",".*-n")] []*[ws(".*-n","\"%w\" has part...",".*-n")] | [ws(".*-n","\"%w\" has part...",".*-
n")] []*[ws(".*-n","\"%w\" is a type of...",".*-n")] []*[ws(".*-n","\"%w\" is the cause of...",".*-n")] |
[ws(".*-n","\"%w\" has part...",".*-n")] []*[ws(".*-n","\"%w\" is the cause of...",".*-n")] []*[ws(".*-
n","\"%w\" is a type of...",".*-n")] |[ws(".*-n","\"%w\" is the cause of...",".*-n")] []*[ws(".*-n","\"%w\"
is a type of...",".*-n")] []*[ws(".*-n","\"%w\" has part...",".*-n")] | [ws(".*-n","\"%w\" is the cause
of...",".*-n")] []*[ws(".*-n","\"%w\" has part...",".*-n")] []*[ws(".*-n","\"%w\" is a type of...",".*-n")])
within <s/>
```

Figure 4. CQL query for the extraction of high-density KRCs

This query reuses the annotations of three types of semantic word sketches (e.g. [ws(".*-n","\"%w\" is a type of...",".*-n")] means any item annotated as a hyponym) and combines them with any other item or items ([]*) that may appear within the same sentence (within <s/>). In this way, we first collected a sample of high-density KRCs that activate at least three different relations: hyponymy, meronymy and causality. The selection of these

relations is based on the grammar rules that are more extensively developed so far in the ESSG.

High-density KRCs in Terminology can be assimilated to what "good dictionary examples" have traditionally been in Lexicography. According to Kilgarriff et al. (2008), a good example must be: (1) typical, exhibiting frequent and well-dispersed patterns of usage; (2) informative, helping to elucidate the definition; and (3) intelligible to learners, avoiding high sentence length, difficult lexis and structures, puzzling or distracting names, anaphoric references, etc.

These parameters were outlined with a collocation dictionary in mind. In our case, typicality is based on the kind of concepts related in the KRCs, depending on whether they refer to typical environmental entities and/or processes or not; informativity depends on the number of conceptual propositions and different relations conveyed in the same KRC; and intelligibility is understood in the same way, although the complexity that can be expected in a domain-specific resource is inevitably higher than in a learner environment. In fact, after performing the query, we manually filtered out all KRCs that: (1) were too long for the scarce amount of knowledge obtained proportionally (see for example Table 4); (2) showed anaphora, as in Table 5, where the particle "such" indicates that not all soils contain calcium ions; (3) contained too many named entities, thus not relating proper concepts (Table 6); or (4) were not actual KRCs because KPs were noisy, as in Table 7, where *part of* does not convey actual meronymy ("another part of the problem"); or where *consist of* is not used as a meronymic KP but rather as a hyponymic one.

They may also authorise, specifying the conditions for: -injections of **water** *containing* **substances** *resulting from* the operations for **exploration and extraction of hydrocarbons or mining activities**, and injection of water for technical reasons, into geological formations from which **hydrocarbons** *or other* **substances** have been extracted or into geological formations which for natural reasons are permanently unsuitable for other purposes.

Table 4. Filtered KRC: high length and low content

Such **soils** are usually *rich in* **calcium ions**, and are often *derived from* **rocks** *such as* **limestone** or **chalk**, which are mostly *composed of* **calcium carbonate**.

Table 5. KRC with anaphora (*such soils*)

Flooding *caused by* **Katrina** was particularly severe *in* the **polder** that *comprises part of* **St. Bernard and Orleans Parishes in New Orleans**, Louisiana, *referred to here as* the **St. Bernard Polder**.

Table 6. KRC containing too many named entities

Another *part of* the **problem** was the **difficulty** in separating anthropogenic changes from natural fluctuations in fish availability, and in **phenomena** *such as* **coastal erosion** and **flooding** which could be *caused by* **human interference or natural processes**.
Managed retreat schemes will *result in* **loss of land** which in many cases may *consist of* **grazing** or other **agricultural land**.

Table 7. Noisy KPs

Finally, approximately 40 high-density KRCs were collected (examples are shown in Tables 8-15). The number of these KRCs involving hyponymy, meronymy and causality in the EEC is relatively low, considering the size of the corpus. This may result from the fact that the ESSG still needs to be refined in order to accommodate new KPs, or the fact that the combination of hyponymy, meronymy, and causality within the same sentence may have a constraining effect on the concepts related and not be as frequent as initially thought. This would actually explain the concept types involved in the vast majority of the KRCs analyzed, which were chemical substances and matter in the case of entities, and formation and transformation processes in the case of events.

The low frequency may also convey that this combination of conceptual relations, found simultaneously in the same sentence, is atypical in the domain. These would therefore be less useful as user-aimed examples in a TKB. They would, however, be useful for a terminographer who wants to build the underlying conceptual system.

Other features that characterize high-density KRCs are the number of conceptual propositions involved and the depth of the hierarchy. In the KRCs analyzed, the number of propositions goes from 3 (as required by the starting query) to 11; and hierarchy depth goes from 1 to a maximum of 3 hierarchical nodes.

For the sake of clarity in our description and characterization of high-density KRCs, we have included numerous examples extracted from the corpus. Tables 8-19, 23, and 26-31 show how KRCs are dissected according to conceptual propositions and hierarchical levels. This qualitative analysis was carried out manually. For instance, the KRC in Table 8 activates 3 propositions and only one hierarchical level, since all other concepts are related to the same starting concept (*fire plume*).

Fire plumes *and other* **biomass fires** *contain* **CO** and *can contribute to* **O3 formation**.		
fire plume	*type_of*	biomass fire
	made_of	CO
	cause	O3 formation

Table 8. High-density KRC: 3 propositions and 1 hierarchical level

In contrast, the KRC in Table 9 shows up to 10 propositions and 3 hierarchical levels. *Photochemical smog* is surrounded by its different components, and all these materials are also related to what they are, what they are made of and what they cause.

Photochemical smog *contains* compounds *such as* aldehydes (*compounds containing* the **-COOH group** *joined directly to* **another carbon atom**), ketones (*compounds containing* the **CO.CO.C group**), and formaldehyde (or methanol, HCHO), which impart a characteristic odor, and nitrogen dioxide (NO2) and solid particles that *cause* a brownish haze.

photochemical smog	*made_of*	compound	*generic_of*	aldehyde	*made_of*	-COOH group
						carbon atom
				ketone		C.CO.C group
				formaldehyde		
				nitrogen dioxide		
				solid particles	*cause*	brownish haze

Table 9. High-density KRC: 10 propositions and 3 hierarchical levels

This does not necessarily mean that the number of propositions correlates with the number of hierarchical levels, since there are KRCs where only 3 propositions may activate 3 hierarchical levels, as shown in Table 10. This builds a conceptual chain starting from a single concept.

Phytoplankton *comprises* at least four thousand **species of plants** that, as on land, use sunlight in the process of photosynthesis to *generate* **sugars** *and other* **high-energy organic compounds**.

phytoplankton	*made_of*	plant	*causes*	sugar	*type_of*	high-energy organic compound

Table 10. High-density KRC: 3 propositions and 3 hierarchical levels

In some of the KRCs analyzed, we found other relations not explicitly searched for, such as *affects* (Table 11) – not as yet formalized in the ESSG but included as a conceptual relation in the environmental TKB EcoLexicon – or *has_location* (Table 12 and 13) and *has_function* – both formalized within the ESSG. This means that in a single sentence we can find even more than 3 different relation types.

In addition, **non-CO2 climate forcers** *(defined as* any **gaseous or particulate compound** that *contributes to* **climate change** *including* **O3**, **CH4**, **nitrous oxide**, **F-gases** *(gases containing* **fluorine**) *as well as* **PM**) *exert influence on* the **Earth's energy balance** and on **climate**.

non-CO2 climate force	*type_of*	gaseous compound		
	causes	climate change		
	generic_of	O3		
		CH4		
		nitrous oxide		
		PM		
		F-gas	*made_of*	fluorine
	affects	Earth's energy balance		
		climate		

Table 11. High-density KRC: other relations (*affects*)

PM is emitted from many sources, and *is a* complex heterogeneous *mixture comprising* both **primary and secondary PM**; primary PM *is the fraction of* PM that is emitted directly into the atmosphere, whereas **secondary PM** *forms in* the **atmosphere** *following* the **oxidation and transformation of precursor gases** *(mainly* **SOX**, **NOX**, **NH3** and some **volatile organic compounds** (VOCs)).

PM	*made_of*	primary PM		
		secondary PM	*has_location*	atmosphere
			caused_by	oxidation and transformation of precursor gases
precursor gas	*generic_of*	SOX		
		NOX		
		NH3		
		VOCs		

Table 12. High-density KRC: other relations (*has_location*)

Ozone *is a* **secondary pollutant** *formed in* the **troposphere**, the lower *part of* the **atmosphere**, *from* complex **chemical reactions** *following* emissions of **precursor gases** *such as* **NOX** and **non-methane VOC** (NMVOC).

ozone	*type_of*	secondary pollutant				
	has_location	troposphere	*part_of*	atmosphere		
	caused_by	chemical reactions	*caused_by*	precursor gas	*generic_of*	NOX
						non-methane VOC

Table 13. High-density KRC: other relations (*has_location*)

The *affects* relation in Table 11 is conveyed through the KP *exert influence on*, which is a KP that has not yet been formalized in the ESSG. This is an example of how the analysis of KRCs can help to extract new KPs to improve the ESSG through the refinement of grammar rules or the formalization of new relations. In the case of the *has_location* relation in Tables

12 and 13, it is conveyed through the KP *form(ed) in*, which is already stored in the ESSG; but in other KRCs, new location-related KPs have been found that might be integrated into the grammar, as is the case with the example in Table 14 (*are prevalent in*).

Still another *constituent of* the **clay traction** *is a group of* **hydrous oxides** *of* **iron** and **aluminum** *known as* **sesquioxides**, which *are prevalent mainly in* the **soils of tropical and subtropical regions** and *are responsible for* the predominantly **reddish or yellowish hue** of these soils.

clay traction	*made_of*	sesquioxide	*type_of*	hydrous oxide
			made_of	iron
				aluminum
			has_location	tropical region
				subtropical region
			causes	reddish or yellowish hue

Table 14. High-density KRC: other relations (*has_location*)

Nevertheless, we must be cautious when integrating new KPs in the ESSG, since they can be more noisy than productive. In this sense, *following* (Tables 12 and 13) can be polysemic, since it can point to a causal relation, as in this case, or a time-related one.

In the same way, prepositions such as *in*, *from* or *for* are very ambiguous. *In* may express location, time, manner, etc.; *from* may convey function, direction, duration, etc.; and *of* may mean part, time, possession, etc. For this reason, they are never included as part of grammar rules unless they are accompanied by a disambiguating element (as in *form(ed) in*), even though prepositions are some of the most important indicators of event-knowledge (Barrière, 2004). However, when encountered within high-density KRCs, their ambiguity can be reduced based on the type of relations and concepts activated around them.

Besides prepositions, there are also equally noisy verbs, such as *have*, which can express meronymy (as in Table 15) but should only be employed as a KP in very controlled scenarios, where its ambiguity can be reduced based on the relations or concepts types activated nearby.

Venus *has* **clouds** *containing* **sulfuric acid** and an **atmosphere** *of* **carbon dioxide** that *produces* a strong **greenhouse effect**, the **Earth** *is* the only **planet** that *has* all three forms of **water** on its surface.

Venus	*made_of*	cloud	*made_of*	sulfuric acid		
		atmosphere		carbon dioxide	*causes*	greenhouse effect
Earth	*type_of*	planet				
	made_of	water				

Table 15. High-density KRC: polysemic verb

Since the number of high-density KRCs collected at this stage was relatively low probably due to the number of relations searched for, the next logical step was to extract high-density KRCs where only two relation types were found. These KRCs are much more numerous in

the corpus and their combinatorial patterns provide further insights into how these KRCs behave.

The results obtained amounted to: 1,392 KRCs showing the combination of *type_of* and *part_of* relations; 3,057 for *type_of* and *caused_by* relations; and 759 for *part_of* and *caused_by* relations. Again, some manual filtering had to be done before analyzing the KRCs, since the same problems reported for the KRCs above were found (e.g. noise, anaphora, KP polysemy, etc.). Then, a random selection was again manually analyzed from a qualitative perspective, since the analysis of the whole set goes beyond the scope and extension of this paper.

Regarding concept types, again chemical substances and matter are the concepts that appear to be most often simultaneously linked through hyponymy and meronymy, although organisms and landforms are also prevalent. In some of the *type_of +part_of* KRCs other relations can be found, such as *affects, has_location, delimited_by* or *has_function*, but in most of them only *type_of* and *part_of* appear. In this case, KRCs tend to show deeper hierarchical levels, as shown in Table 16, where a fourth hierarchical level appears.

The typical **aluminosilicate clay minerals** appear as laminated microcrystals, *composed mainly of* two basic **structural units**: a **tetrahedron** *of* four **oxygen atoms** surrounding a central **cation**, *usually* **Si4+,** and an **octahedron** *of* six **oxygen atoms** surrounding a somewhat larger **cation** of lesser valency, *usually* **Al3+** or **Mg2+.**							
Al. clay mineral	*has_part*	structural unit	*has_type*	tetrahedron	*made_of*	oxygen atom	Si4+
				ocahedron		cation	Al2+
						type_of	Mg2+

Table 16. High-density KRC: 8 propositions and 4 hierarchical levels (type_of+part_of)

Type+cause KRCs are the most numerous of the two-relation combinations, prototypically activating both processes and entities related to organisms, landforms, coastal structures, substances, wave action, meteorological phenomena, and chemical reactions and diseases, which is only natural, since causality involves the description of processes and their participants. For example, Tables 17-19 show three KRCs where only *type_of* and *caused_by* are activated.

Non-electrical energy sources, *such as* **natural gas furnaces**, also *produce* **GHGs**.				
non-electrical energy source	*has_type*	natural gas furnace	*causes*	GHG

Table 17. High-density KRC: 2 propositions and 2 hierarchical levels (type_of+cause)

Water (mainly **acidic water**) and **gases** in air or dissolved in water (mainly **oxygen** and **carbon dioxide**) can cause **chemical weathering**.				
chemical weathering		acidic water	*type_of*	water
	caused_by	oxygen		gas
		carbon dioxide		

Table 18. High-density KRC: 5 propositions and 2 hierarchical levels (type_of+cause)

Climatic hazards such as **strong winds** and **heavy rains**, **storms** and **hurricanes** frequently give rise to **landslides**, **sediment flows**, and **water floods**.				
strong wind				landslide
heavy rain	*type_of*	climatic hazard	*causes*	sediment flow
storm				water flood
hurricane				

Table 19. High-density KRC: 7 propositions and 2 hierarchical levels (type_of+cause)

In contrast to *type_of+part_of* combinations, these KRCs seem to reflect less deep hierarchical levels (i.e. only two in the previous examples), no matter the number of conceptual propositions reflected (e.g. from 2 to 7 in the previous examples), at least when only *type_of* and *caused_by* are activated. However, causality and hyponymy are found very often in combination with other relations not yet developed in the ESSG, such as *affects*. Only in these cases deeper hierarchical levels emerge. Furthermore, most KRCs of this kind are not as straightforward as the ones shown in Tables 17-19. The expression of causality often includes entire clauses that are difficult to merge in a single concept, as in Tables 20 and 21, where "variations in the force of gravity", "relative movements of heavenly bodies", "the moon revolving around the Earth", "movements of marine boundaries", etc. would hinder the construction of concept systems. Therefore, despite the fact that these contexts contain unambiguous KPs pointing to useful semantic relations, their use as KRCs may be restricted to terminology users rather than terminographers when building a TKB.

The **astronomical tide** refers to the **regular oscillations of the sea or ocean surface**, due to **variations in the force of gravity** caused by the **relative movements of heavenly bodies**, mainly **those of the moon revolving around the Earth, and those of the Earth, revolving around the Sun**.

Table 20. High-density KRC: clauses

Tidal waves or **tsunamis** are **long waves**, generated by **movements of marine boundaries**, such as the **sea bottom** because of a **submarine earthquake**, or of a **sea slope** because of a **sediment slump** or **landslide** above or underneath the water.

Table 21. Table 20. High-density KRC: clauses

As for *part_of+caused_by* KRCs (Tables 22 and 23), the concepts involved are related to the parts and materials of the same concept types related in *type_of+caused_by* KRCs in

addition to other categories of concepts such as tools and instruments or physical and mathematical concepts. Again in the case of this type of KRCs, the hierarchical depth seems to be lower except when other relations apply (Table 23; *affects*).

Ground-level O3 and **black carbon**, *a constituent of* **PM**, *contribute to* **global warming**.				
particulate matter	*has_part*	black carbon	*causes*	global warming
		ground-level O3		

Table 22. High-density KRC: 4 propositions and 2 hierarchical levels (part_of+cause)

In temperate climates, **beaches** typically *consist of* **quartz** and **feldspar grains** *derived from* the **weathering** *of* **terrestrial rocks**.						
beach	*has_part*	quartz grain	*caused by*	weathering	*affects*	terrestrial
		feldspar				rock
		grain				

Table 23. High-density KRC: 5 propositions and 3 hierarchical levels (part_of+cause)

In this case it is also very common to find causal clauses, as shown in Table 24 (e.g. "exposure of GAC surface to such effluents").

Since the **effluent** of primary treatment *contains* a relatively high amount of **biodegradable substances**, the **exposure of GAC surface to such effluents** *leads to* rapid **colonization by microorganisms**.

Table 24. High-density KRC: clauses

Finally, since causality seemed to be widely represented in the corpus (even though causal grammars are not as developed as others, such as hyponymy), a different query was performed in order to extract high-density KRCs where different propositions of the same kind (i.e. causality) could be extracted. In this case the analysis was more focused on the type, amount and diversity of causal KPs involved in these KRCs (i.e. *cause, caused by, produce, generated by, give rise to, derived from, trigger*, etc.). The initial set of the extracted contexts amounted to more than 7,000. After a manual filtering and a random selection, a set of 260 was retained and classified according to the KPs contained. Most of them only contained two KPs, but 38 of them contained 3 or more KPs. The maximum number of KPs encountered in the same KRC was 7, as shown in the KRC in Table 25. However, again in this case not all propositions could be used in a concept system due to the nodes expressed in the form of clauses instead of self-contained concepts.

There are four primary meteorologic causes that should be recognized and the data segregated accordingly: (1) **thunderstorm type events** where the *resulting* **flood** is *caused by* high intensity, short duration, **rainfall** that *produces* **high peak discharges** and relatively low volumes (2) **general rain type events** where the *resulting* **flood** *is caused by* moderate intensity, long duration, **rainfall** (3) **snowmelt floods** *resulting from* the **melting of an accumulated snow pack** and (4) **floods** *resulting from* a combination of **rain falling on a melting snowpack**.

Table 25. High-density KRC: causality through 7 KPs

Some of these KRCs include several conceptual propositions through the same KP, such as the one shown in Tables 26 (i.e. *produce*), 27 (i.e. *cause*), 28 (i.e. *produce*) and 29 (i.e. *induced by*). It usually occurs in enumerations (i.e. Tables 26 and 27), when the causes and/or effects belong to the same conceptual category, especially co-hyponyms (i.e. Table 28), or when the same cause or effect participates in several propositions, usually acting as the subject or object of the sentence (i.e. Table 29).

Typical process-form inferences include: **creep processes**, which *produce* an expanding **upper convexity** on a slope; **erosion by overland flow**, especially by gullying, which *produces* an increasing **lower concavity**; **uniform solution**, which *produces* a parallel **downwearing**; and shallow **landslides**, which *produce* parallel **slope retreat**, associated with a lower concavity in which the slide debris accumulates.

creep process		upper convexity
erosion by overland flow	*causes*	lower concavity
uniform solution		downwearing
landslide		slope retreat

Table 26. High-density KRC: 4 propositions and 1 hierarchical level (cause+cause)

However , **sulfates** *cause* other **environmental problems** *such as* **acid rain**, **carbon black** *causes* **human health problems**, and **dimming** *causes* **ecological problems** *such as* **changes in evaporation and rainfall patterns**, with **droughts** and **increased rainfall** both *causing* **problems for agriculture**.

environmental problem	*has_type*	acid rain	*caused_by*	sulfate
carbon black	*causes*	human health problem		
dimming	*affects*	evaporation rainfall	*affects*	Agriculture
drought				

Table 27. High-density KRC: 8 propositions and 2 hierarchical levels (cause+cause)

Steep pressure gradients (tightly packed isobars) *produce* **strong pressure gradient forces** and **high winds**; **gentle pressure gradients** (widely spaced isobars) *produce* **weak pressure gradient forces** and **light winds**.

steep pressure gradient		strong pressure gradient force
	causes	high wind
gentle pressure gradient		weak pressure gradient force
		light wind

Table 28. High-density KRC: 4 propositions and 1 hierarchical level (cause+cause)

Water level changes can be locally *induced by* **winds** blowing across the bay or they can be *induced by* **flow of water** through inlets from surges generated on the open coast.		
water level change	*caused by*	wind
		flow of water

Table 29. High-density KRC: 3 propositions and 1 hierarchical level (cause+cause)

In constrast, there are other KRCs where causality is expressed with different KPs each time. In these cases the number of conceptual propositions and the complexity of causality tend to increase (Table 30).

Along the studied beaches, the persistent **longshore transport gradient** can *result from* **longshore transport divergence** *induced by* **wave refraction** over an ebb-tidal shoal, **flood currents** along the beach proximal to a tidal inlet, **increased wave energy** *due to* a nearshore **dredge pit**, and total **littoral blockage** *by* structured inlet with a minimal ebb-tidal shoal.				
longshore transport gradient	*caused_by*	longshore transport divergence	*caused_by*	wave refraction
		flood current		
		increased wave energy	*caused_by*	dredge pit
		littoral blockage	*caused_by*	structured inlet

Table 30. High-density KRC: 3 propositions and 1 hierarchical level (cause+cause)

In the case shown in Table 27, despite the KP repeatedly used (i.e. *cause*), not only causality is actually codified, since some of the effects can be understood in terms of the *affect* semantic relation. For example, the clause "causing problems for agriculture" should be understood as "something *affects* agriculture", since "problems for agriculture" is not a concept itself. Also, "causes ecological problems such as changes in evaporation and rainfall" can be conceptually rephrased as "something *affects* evaporation and rainfall".

As can be observed, in causal KRCs, multiple conceptual propositions may be extracted but few hierarchical levels are activated by most of them. We can encounter KRCs showing several different causes linked to their effects through a single hierarchical level (e.g. Table 26 or 28) or a central concept from which a causal chain is deployed (e.g. Table 31), but not more than 2 levels have been found. In the first case, there will be more chances of finding the same KP conveying causality, whereas in the second, causality will be conveyed through a greater variety of KPs.

These types of **earthquakes** also frequently *cause* large **submarine** (underwater) **landslides** or **slumps**, which also *generate* **tsunamis**.				
earthquake	*causes*	submarine landslide	*causes*	tsunami
		submarine slump		

Table 31. High-density KRC: 4 propositions 2 hierarchical levels (cause+cause)

Conclusions and Future Work

In this paper we have shown how the combination of word sketches with KPs can provide a reliable user-friendly method for the extraction of KRCs. KRC extraction can be further improved by agglutinating different semantic word sketches in the same query, thus giving rise to high-density KRCs. Three different sets of this kind of KRCs were collected: (1) KRCs where three different semantic relations could be found; (2) KRCs codifying two semantic relation types but several conceptual propositions; and (3) KRCs codifying only a semantic relation type but several conceptual propositions.

As the number of KRCs with three different semantic relations (hyponymy, meronymy and causality) was relatively low, the next logical step was to extract high-density KRCs where only one or two relation types were found. These KRCs were much more numerous in the corpus and their combinatorial patterns provided further insights into how these KRCs behave.

High-density KRCs were characterized based on the amount of knowledge conveyed, the concept types involved, the conceptual depth activated and the number and type of conceptual relations codified in them. For instance, in KRCs combining *type_of* and *part_of* relations, deeper hierarchical structures can be found compared to the rest.

Not all the high-density KRCs collected were equally useful for different types of users. For example, KRCs that express causality often included entire clauses that are difficult to merge in a single concept. Therefore, despite the fact that these contexts contained unambiguous KPs pointing to useful conceptual relations, their use as KRCs may be restricted to terminology users rather than terminographers when building a TKB. In contrast, KRCs where too many conceptual propositions are activated may be more useful for terminographers than for terminology users.

The analysis of KRCs can help to extract new KPs to improve the ESSG through the refinement of grammar rules or the formalization of new relations. Nevertheless, we must be cautious when integrating new KPs in the ESSG, since they can be more noisy than productive, as is the case of prepositions such as *in* and *from*, and verbs such as *have*.

As future work, we plan to expand this analysis to KRCs beyond one sentence and to KRCs where hyponymy and meronymy are combined with the *has_location* or *has_function* relations, which are already developed in the ESSG. A more thorough and quantitative analysis will be conducted on those KRCs where only two relations apply but from which several propositions may be extracted.

Furthermore, the extraction of high-density KRCs will be more productive as soon as the ESSG is refined and improved. Finally, following the approach of Condamines et al. (2013), it would be interesting to explore the usefulness of high-density KRCs as well as the cognitive effort they require for end users such as translators.

Notes

1. The EEC is freely available in the.sketchengine.co.uk/open and the ESSG can be downloaded and reused with any other corpus following the instructions on http://ecolexicon.ugr.es/essg
2. More details on the Dublin Core Metadata Initiative can be found on http://dublincore.org/specifications/dublin-core/
3. The latest version of the ESSG can be downloaded from http://ecolexicon.ugr.es/essg/
4. By conceptual proposition we mean the triple concept-relation-concept. QUARTZ *type_of* MINERAL would be a hyponymic proposition.

References

Aussenac-Gilles, N., & Jacques, M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with Caméléon. *Terminology, 14*(1), 45-73. Retrieved from http://doi.org/10.1075/term.14.1.04aus

Barrière, C. (2004). Knowledge-Rich Contexts Discovery. In *Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)*, Vol. 3060. London, Ontario: CSCSI, 187-201. Retrieved from http://doi.org/10.1007/978-3-540-24840-8_14

Barrière, C., & Agbago, A. (2006). TerminoWeb: a software environment for term study in rich contexts. In *Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006)*. Beijing.

Bowker, L. (2003). Lexical Knowledge Patterns, Semantic Relations, and Language Varieties: Exploring the Possibilities for Refining Information Retrieval in an International Context. *Cataloging & Classification Quarterly, 37*(1–2), 153-171.

Condamines, A., Josselin-Leray, A., Fabre, C., Lefeuvre, L., Picton, A., & Rebeyrolle, J. (2013). Using Comparable Corpora to Characterize Knowledge-Rich Contexts for Various Kinds of Users: Preliminary Steps. In *Proceedings of the 5th International Conference on Corpus Linguistics. Procedia-Social and Behavioral Sciences 95*, 581-586.

Faber, P., León-Araúz, P., & Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era. Educational Linguistics, 19*. Springer, 267-301.

Faber, P., León-Araúz, P., & Reimerink, A. (2016) EcoLexicon: new features and challenges. In I. Kernerman, I. Kosem Trojina, S. Krek, & L. Trap-Jensen (Eds.) *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, Portorož, 73-80.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.) *Proceedings of the Eleventh EURALEX International Congress*. Lorient: EURALEX, 105-116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008) GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (Eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra.

León-Araúz, P., San Martín, A. & Reimerink, A. (2018) The EcoLexicon English Corpus as an open corpus in Sketch Engine. In *Proceedings of the XVIII EURALEX International Congress*. Ljubljana, Slovenia.

León-Araúz, P. & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In I. Kerneman, & S. Krek (Eds.) *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*. Miyazaki: Globalex, 94-99.

León-Araúz, P., San Martín, A. & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology*. Osaka, Japan, 73-82.

Marshman, E. (2002). The cause-effect relation in a biopharmaceutical corpus: English knowledge patterns. In *Proceedings of the 6th international Conference on Terminology and Knowledge Engineering*. Nancy, 89-94.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, M.C. L'Homme, (Eds.) *Recent Advances in Computational Terminology*. John Benjamins, Amsterdam, 279-302.

San Martín, A., Cabezas-García, M., Buendía, M., Sánchez-Cárdenas, B., León-Araúz, P. & Faber, P. (2017) Recent Advances in EcoLexicon. *Dictionaries: Journal of the Dictionary Society of North America*, 38(1):96-115.

Grammatical metaphor at tertiary level: Rise, development, and implications revisited

Eka Fadilah[*] - Universitas Widya Kartika Surabaya, Indonesia
Mirjam Anugerahwati - Universitas Negeri Malang, Indonesia

ABSTRACT

This article argues for the need to maintain grammatical metaphor as a powerful tool for meaning construal that is applicable in both written and spoken languages, notably at the tertiary level. The features in grammatical metaphor enable the writer and speaker to use language based on its use, form, and function. Given the uniqueness of grammatical metaphor constructions compared to traditional grammar approach, it is necessary to revisit its concept, application, and implication to illuminate some pedagogical and empirical impacts on language teaching and learning. Furthermore, some recommendations to incorporate grammatical metaphor into language curriculum are proposed by taking the proliferation of ubiquitous information, communication and technology as a source to create a new genre.

Keywords: grammatical metaphor; rise; development; implication; tertiary level

RESUMEN

Este artículo argumenta por la necesidad de mantener a la metáfora gramatical como una herramienta para la construcción de significado aplicable en lenguas escritas y orales, notablemente en el nivel terciario. Las características de la metáfora gramatical permiten a escritores y lectores emplear la lengua basada en su uso, forma, y función. Dada la originalidad de las construcciones metafóricas gramaticales en comparación con el enfoque gramatical tradicional, es necesario revisitar este concepto, su aplicación, y sus implicancias para iluminar algunos impactos pedagógicos y empíricos en la enseñanza y aprendizaje de la lengua. Además, recomendaciones para la incorporación de la metáfora gramatical son propuestas tomando la proliferación de las tecnologías de la información como fuente creacional de un nuevo género.

Palabras clave: metáfora gramatical, elevación, desarrollo, implicancias, nivel terciario

---

* Corresponding author, e-mail: mref3k4@gmail.com

THERE HAS BEEN a great deal of discussion in reference to grammatical metaphor, henceforth GM, since Halliday (1985) introduced the idea in *An Introduction to Functional Grammar*. Such discussions encompass, but are not limited to, definition, classification, function, realization, forms, and working mechanism of GM (e.g., Christie & Derewianka, 2008; Halliday, 1985; Halliday & Matthiessen, 2004, 2014; Halliday & Webster, 2009). Additionally, GM is reported as a part of English teaching and learning (Table 1).

GM is rooted in Halliday's (1985) Systemic Functional Linguistics (SFL) in which language is conceived as a semiotic system that is mapped on strata encompassing expression and content. Such strata are interrelated, for example content layer embodies lexicogrammar (wordings) and semantic discourse (meanings), while expression layer constitutes phonology (sounds) (Halliday & Matthiessen, 2014; Martin & Rose, 2008). The stratified system constitutes essential hierarchy of abstraction by recoding the language across the strata (Devrim, 2015), for instance, semantic discourse is realized by lexicogrammar by shift-ranked some elements from lexicogrammar (e.g., noun, verb, adjective, conjunction) to semantic discourse (participant, process, quality, relator). Additionally, as a semiotic system – system of meaning in which "it has the further property that it is a *semogenic system* – a system that creates meaning," but not all semiotic systems constitute semogenesis, such as a system of traffic signals as a system of meaning but they cannot create meanings built into it (Halliday, 2009, p.60). SFL embodies a general inseparable concept of grammar and lexis rather than a triad of syntax, morphology, and lexicon. The rationales of such a concept are explained as (1) the distinction of i.e., syntax and morphology "has always been ill-defined" rather than term "grammar" and (2) the unclear boundary between grammar and lexicon entails a continuum pool of grammar and lexis on a single stratum "lexicogrammar" (Halliday, 2009, p. 73).

Grammatical metaphor

GM is defined as a "tension between lexicogrammar and discourse semantic by people mean language into more than one thing at once, effectively expanding the language's meaning potential" (Martin, 2008, p. 829). According to Halliday, grammar is a resource of meaning-making which enables the work of grammar to unconsciously construe experience (Webster, 2009). Such a construction enables us to either re-theorize or re-make the meaning variation at a different level (Halliday, 2004). As Halliday provides the following example:

1. *(a)* Glass cracks more quickly the harder you press on it.
   (b) The rate of glass crack growth depends on the magnitude of stress.

<div align="right">(Cited in Webster 2009, p.4)</div>

Sentence (1a) illustrates congruent sentence with elements: glass (participant), crack (process), more quickly (circumstances), the harder (circumstances), you (participant), press (process), on (circumstances), and it (participant). While, Sentence (1b) demonstrates the re-

wording or re-making of approximate meaning by projecting a more scientific text, in which words cracks and press (process) are turned into Things (Noun). For Halliday, nominalization that realizes a certain clause in lexicogrammar stratum and semantic discourse stratum is the congruent realization, while the nominalization that has two or more realizations in lexicogrammar stratum and semantic discourse stratum is the incongruent realization. As Halliday (1994, p.343) defined a congruent realization as "the typical ways of saying things," and an incongruent one as "not expressed through the most typical (and highly coded) form of representation" (Halliday, 1978 cited in He & Wen, p.3). Which one is better, congruent or incongruent dimension? Likewise, Christie and Derewianka (2008) argue that GM "serves to organize text and compact information creating high level of lexical density" (p.116). As nominalization, GM enables academic text to be construed not only in technicality but also in rationality by constructing logical semantic relations in the text. Thus, such a nominalization is inevitably needed in a scientific argument as a strand position of a scientist as Halliday proposed:

> The core of scientific text was the development of a chain of reasoning (ultimately based on experiments) in which each step led on to the next. But in order to lead on to the next step, you have to be able to repeat what has gone before and now is being used as a springboard for the next move (cited in Webster, 2009, p.4)

Types of Grammatical Metaphor

Initially, Halliday (1985) comprised GM into two models: *ideational* and *interpersonal metaphors*. Eventually, Martin (1992) proposed *textual* metaphor to be useful term "when discourse systems are used to construe text as 'material' social reality" (p.416). Among the other two types of metaphor, textual metaphor remains an unelaborated and controversial issue in SFL (Halliday & Matthiessen, 2014; Yang, 2018b).

Ideational metaphor is comprised of into two types; *experiential* and *logical* metaphors (Halliday, 1985; Martin, 1992). The former demonstrates the construal meaning of tension between lexicogrammar and discourse semantic where *process* is realized by verb, *quality* is realized by adjective, entity is realized by noun, and adverb or preposition is realized by circumstances. While logical metaphor constitutes construal meaning in the consequential and temporal relations inside clauses (Martin, 1992). These two types of ideational metaphor interact with conjugative relations, which are realized metaphorically and such interaction enables "a high level of abstraction in text, making it inaccessible to large sections of the community" (Martin 1992a, p. 407). This is demonstrated in the following sentences:

2. (a) Because internet **gets better**, Indonesian people are able to communicate with other people around the world **more easily**.

    (b) Internet gets better, so Indonesian people are able to communicate with other people around the world **more easily**.

(c) The **advances** of internet **enable** Indonesian people to global communication **more easily**

(d) The advances of internet **facility** lead Indonesian people to global communication.

Sentences 2 (a) and 2 (b) are congruent forms in which sentence 2 (a) consists of two clauses as *hypotaxis* (unequal status), similarly, sentence 2 (b) comprises two clauses as *parataxis* (equal status) in which both sentences construct some equal elements of language stratifications: lexicogrammar and discourse semantic: Conjunction is equal to relator (because), noun is equal to participant (internet, Indonesian people, other people, the world), verb is equal to process (gets, are able to, communicate), adjective is equal to quality (better), preposition is equal to circumstances (with, around), adverb is equal to circumstance (more easily). These are what is called *congruent* forms.

By contrast, sentences 2(b) and 2(c) denote some realization from one element to another, for instances, *because* (conjunction) is replaced by *enable, lead* (process), *gets better* (adverbial group) is replaced by *advances* (participant), *more easily* (adverbial group) is replaced by *facility* (participant), *are able to* is realized by *enable*, *to communicate* (process) is replaced by *communication* (entity), and *with other people around the world* (prepositional group) is reaized by *global* (quality), these substitutions and transcategorization are called as *incongruent forms*.
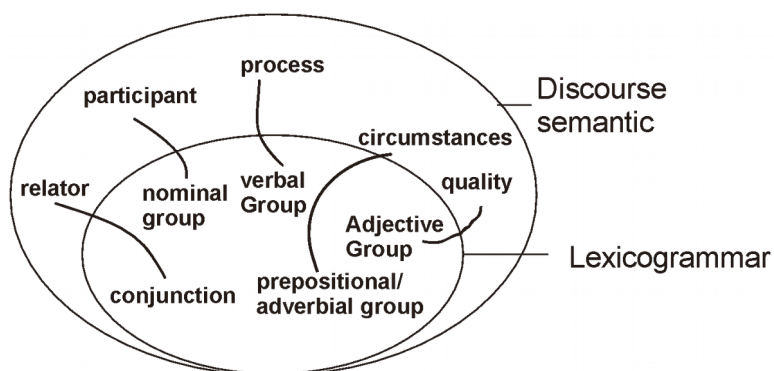

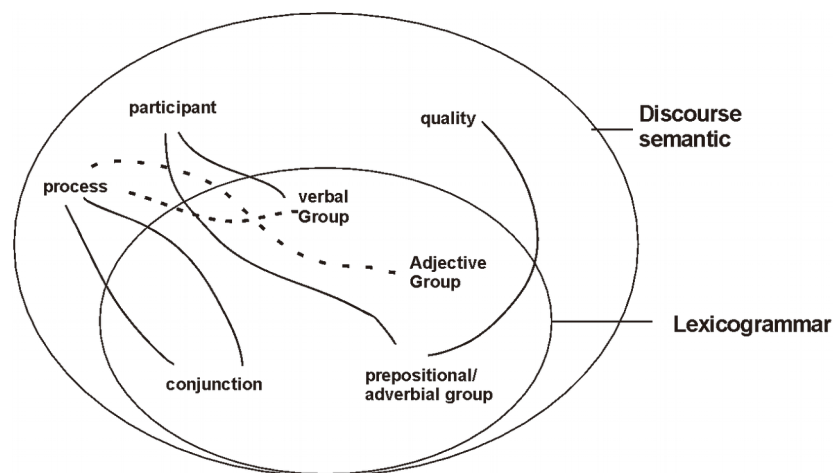Figure 1. Classification of congruent forms.

Figure 2. Substitutions and transcategorization of incongruent form.

Interpersonal Metaphor constitutes linguistic resources to adjust the negotiability and level of controversy of an utterance that is used to "enact social relationships through interactions between speakers/writers and listeners/readers" (Yang, 2018a, p.2). *Interpersonal metaphor* is comprised by metaphor of *mood* and *modality* as a tension between lexicogrammar and discourse semantic. The former construes a discourse semantic as an incongruent mood form in grammar. Metaphor of mood is typically found in spoken language (see e.g., Devrim, 2015; Halliday & Matthiessen, 2014). Halliday and Matthiessen (2014) variables as message exchanges as *give* and *demand* that substitutes *goods-&-services* and *information*. Those variables invoke four primary speech functions that are "offer, command, statement, and question" (p.137). These speech functions "are realized by linguistic forms *declarative, interrogative, and imperative*" (Yang, 2018a, p.4). Thus, when pairing speech functions *statement-declarative, question-interrogative, command imperative, and offer-interrogative or declarative* occurs, it is called a congruent form. By contrast, when inconsistent pairings are found between *command–declarative (3b-3b), statement - interrogative (4a - 4b), command – declarative (5a - 5b), interrogative – declarative (6a - 6b),* it is called as *mood of metaphor*.

3. *(a)* Look at the way they cheated before
   (b) The evidence is (the fact) that they cheated before
4. *(a)* You shouldn't say such a thing
   (b) How could you say such a thing?
5. *(a)* Don't move or I'll shoot
   (b) If you move I'll shoot
6. *(a)* Tentatively is the position still available?
   (b) I was wondering if the position is still available

(Cited in Halliday, 1994, p.366-367)

On the other hand, in metaphor of modality, Halliday (1994) encompasses modality in SFL as probability, use, obligation, and inclination. These types of modality are categorized into four semantic domains: subjective, objective, implicit, and explicit. Example (2) is from Halliday (1994, p. 354) as modality metaphor:

7. *(a)* **probably** that pudding never will be cooked.
   (b) **I don't think** that pudding ever will be cooked.

Direct implicit subjective in Sentence (7a) is reworded as indirect explicit subjective in sentence (7b). It is claimed that *the direct form is congruent, while indirect form is incongruent.*

On the other hand, textual metaphor is conceived to be included in the study of GM. Martin (1992) asserts that "certain discourse elements organize text rather than field, which include meta-message relation, text reference, negotiating text and internal conjunction, all of which are text-organizing pro-forms" (p. 416–417). Some conjugative links and their categories belong to textual metaphor, that is *reason, example, point, factor, pointing out* as meta-message relation, *this* as *text reference, let me begin by* as negotiating texture, and *a number of reasons, for example, another example, as a final point, as a result of these factors* as internal conjunction.

8. I think Governments are necessary at different levels for ***a number of reasons. Let me begin by*** pointing out that the Federal Government fixes up problems that occur in the community. ***Another example*** is that the State Government looks after schools; ***this*** prevents vandalism and fighting. ***As a final point*** the Local Government is important to look after rubbish: otherwise everyone would have diseases. ***As a result of these*** factors, Governments at several administrative levels are necessary. (Martin,1992, p. 416–417).

Example (8) explicates the subsequent use of textual metaphor categories among internal conjunction: *a number of reasons, another example, as a final point and as a result of these factors,* negotiating texture: *let me begin by pointing out that,* and text reference: *this* in a text.

Grammatical Metaphor as Nominalization and Verbalization

Nominalization constitutes the core construct underlying GM rooted in SFL functions in constructing knowledge, enabling evaluation, and facilitating information flow which affects the relationship between discourse semantics and lexicogrammar (see e.g., Byrnes, 2009; Halliday & Matthiessen, 2014; Martin, 2008; Yongsheng, 2008). As Martin (2008) said, the distinctions between lexicogrammar and discourse semantics as "the grammar and semantics match because we have a semantic entity realized by a noun" (p.802). Additionally, Nominalization entails transformation in terms of *derivation and agnation* (Ryshina-

Pankova, 2010) from one element of grammar (e.g., verb, adjective, adverb) to another element discourse semantics (e.g., participant, process, quality; Christie & Martin, 2007). When one element matches the other it is *congruent*, whereas when one element is transformed into one or more elements it is incongruent or (grammatical) metaphor (Halliday, 1985; Halliday & Matthiessen, 2004; 2014).

In a nutshell, as a semiotic system, first, Halliday comprises language meaning-making as – organizing language into stratal models that is phonology, lexicogrammar, and semantic discourse) - called as *stratification* (see figure 1). Second, the interaction between one language stratum with the other(s) signifies *realization* – meaning expression in upper stratum (semantic discourse) is realized by lower stratum (lexicogrammar) (see Figure 2). Third, the realization invokes language metafunctions – various language functions explicating ideational, interpersonal, and textual functions or metaphors – which construe some variations of grammatical and lexical forms as grammatical/lexical metaphor through *transcategorization* – changing one word class by using *derivational* (e.g., *inform* as verb – *information* as noun) or *non-derivational* (e.g., *work* as verb – *work* as noun) morphology as the fourth system. Additionally, transcategorization also involves *agnation* – the relation between congruent and incongruent (metaphorical) construct of clauses, for instance, the incongruent forms *allocation of the entire avenue, alcohol impairment,* and *access to the computer* come from *they allocate the entire avenue, it is impaired alcohol,* and *they were able to reach the computer* (see e.g., Halliday & Matthiessen, 2004, 2014).

Yongsheng (2008) criticizes the overwhelmingly emphasis of GM as nominalization by putting less attention to GM as verbalization form. Verbalization is defined as "the language phenomenon that something that is not a process by itself is taken as a process or a non-action is realized by a verb" (p.301). GM as verbalization encompasses some classifications as a process of *temporal relation, cause-effect relation, condition, and concession.*

Some connective verbs are expressed as *temporal relation* e.g., *while, meanwhile, concurrently, simultaneously, and at the same time* used to express two or more than two events happening at the same times, while *conjunctive expression* e.g., *before, after, afterwards* are used to express two or more than two events happening at the different time.

9. *(a)* There was a strong earthquake in *Central Java*, and **at the mean time** the buildings were devastated. *(congruent)*
   (b) A strong earthquake **was accompanied** by building devastation in Central Java. *(Incongruent)*

10. *(a)* There was a strong earthquake in *Central Java*, **afterwards** the buildings were devastated. *(congruent)*
    (b) A strong earthquake **followed** building devastation in Central Java. *(Incongruent)*

As cause-effect relation, verbalization may use some linking markers to express the reasons such as because, for, since, therefore, so, etc. While as a process of condition, verbalization can use connective links: and, unless, if, etc., eventually, as concession process, verbalization may use some conjugative expressions: although, even though, nevertheless, etc.,

11. *(a)* Because the oil price increases, the government warns to use oil efficiently. *(congruent)*

(b) The increase of oil price leads to warning of efficient oil use. *(incongruent)*

12. *(a)* If the oil price increases, laborers will strike to against it. *(congruent)*

(b) The strike of laborers is determined by the increase of oil price. *(incongruent)*

13. *(a)* Although the oil price increases, the laborers are silent. *(congruent)*

(b) The silence of laborers doesn't mean their agreement of the oil price increase. *(incongruent)*

There are some similarities and differences with reference to GM as nominalization and verbalization. The similarities refer to the constructions that (1) both have the tension between lexicogrammar and semantic discourses, (2) both can make re-construction of meanings in different ways, and (3) both may interface one to another that is "the use of nominalizations lead to the occurrence of verbalization and *vice versa*" (Yongsheng, 2008, p.305).

Pedagogical and Empirical Implication of GM

Theoretical and empirical groundwork of GM aforementioned have explicated and reported how GM constructs text, clause, and discourse embodying foregrounded links between lexicogrammar and semantic discourse. It is necessarily to say that analysis of learners' written and spoken language are not merely skewed on the analysis of genre types (e.g., recount vs. argumentative), grammatical errors (e.g., s-v agreement, article use, etc.,), but also the broader types subject matter (field) and audience type (tenor). GM enables to encode data derived from those broader types to be discussed and specified thoroughly (Rishyna-Pankova, 2015).

Academic writing, for instance, necessitates expert writers i.e., lecturers to be appraised for their professionalism to publish or perish as well as the novice writers i.e., learners to write their final projects e.g., research projects, theses, dissertation, requires those writers to provide a condensed text elucidating the major issues raised in the research report. One of the systemic strategies deals with such academic writing is applying GM and any genre e.g., recount, description, argumentation, exposition that could be used in writing research subchapters e.g., Introduction, method, discussion. Such strategies are conceived as powerful tools in writing various scientific and technical texts (Devrim, 2015; Halliday & Matthiessen, 2014; Webster, 2009).

GM is conceived to not only equip learners with the experience of understanding meaning metaphorically, but also guiding them to evaluate, refine, and redefine the quality of the construal meaning (Liardet, 2016). Such construal can be used as text cohesion, language condensation, lexical density, and logical reasoning (Byrnes, 2009; Liardet, 2013; Ryshina-Pankova, 2010). The use of nominalization construed in GM leads to the way to express objectification and abstraction that are mostly used in scientific texts and academic registers (Ryshina-Pankova, 2010). Furthermore, Liardet (2013) highlights the efficacy of GM as (1) anaphoric reconstrual (building an argument by summarizing and restating what was previously stated), (2) elaborated nominal group (infusing multiple meanings into a single clause element), (3) networking cause and effect (interaction of logical and experiential meanings), and (4) meaning accumulation through (repetition or restatement). Similarly, GM contributes to the higher level of writing abstraction such as the use of GM to paraphrase or summarize an academic text (Yasuda, 2017). For example, Yasuda reports that GM promotes learners to write more rhetorically effective and coherently structured in writing summary. It also leads the learners to think the authors' stance in the original texts as implied statement rather than direct statement by using nominalization (GM).

Ryshina-Pankova and Byrnes (2013, p.195) coined "thinking of writing" which includes activities that teacher-learners do in discussing GM in the subsequent steps such as: (1) asking the learners to highlight GM as nominalization and its roles in the texts, (2) demanding the learners to highlight GM types: ideational, functional, and textual functions and how they construe that way either in academic texts or spoken registers, and (3) asking the learners to unfold and unpack GM by providing the congruent forms to be shifted into incongruent one by paraphrasing, summarizing, synthesizing, etc., Indeed, such processes cannot be conducted at once time in one occasion, but rather the GM configuration and reconfiguration need to consider learners' level of proficiency. For instance, the first step might fit a beginning level by discussing the process of word formation (e.g., derivation, agnation) to enhance learners' vocabulary and grammar. The second step might be better suited for intermediate level by encouraging learners to shift from personal experience to more thematic areas such as particular foreign culture. The third step could be implemented at the advanced level after having experienced with the foregoing two steps. Yasuda's (2015) finding supports such idea in which GM is influenced by learners' level of English proficiency. The higher level of proficiency, the more metaphorical clauses use compared to the lower proficiency learners.

Furthermore, the considerable research findings on GM enable researchers to define and re-define the GM implementation. Devrim (2015) put forward that researchers have three options to investigate the GM constructs that is "following the *stratal model,* adopting the *semantic model,* or theorizing their own" (p.13). The two former options can be conducted by replicating the previous research findings by investigating the use of GM in learners' written or spoken language (e.g., Liardet, 2013, 2016; Yasuda, 2015, 2017),

analyzing academic texts or article abstracts (Ryshina-Pankova, 2010). While the latter option, researchers might start by analyzing language registers through Corpus study (e.g., He & Yang, 2018; He& Wen, 2017; Hu, 2015; Liardet, 2018). Some corpora studies e.g., Corpus of Historical American English (COHA), Corpus of Contemporary American English (COCA), British National Corpus (BNC) are the largest corpus containing larger sizes with 400, 520, and 100 million words respectively (He & Wen, 2017). Besides, the researchers might also create their own corpora study by compiling either learners' written or spoken language. A contrastive or comparative study of English use between outer and expanding countries could provide a clear picture with regard to lexicogrammatical variations used in variety of genres (Hu, 2015). However, it should not merely analyze the concurrent frequency of GM appearing in the texts across genre, but rather the analysis that accounts for the nuances "greater variation, quality and effect of its deployment," and it might be achieved by providing instructional support and understanding of such nuances (Liardet, 2016, p.117).

Recommendation and Conclusion

GM constitutes a powerful tool for language and content integration in language pedagogy as a part of *genre* analysis, especially at tertiary level. Indeed, grammar based-syllabus is still widely practiced in Teaching English as a Foreign Language (EFL), however, such a practice is by no means of regressing to the traditional method (e.g., Audio-Lingual Method) but rather to adapt to fit the context (Fadilah, 2018). Likewise, the ubiquity of internet and proliferation of digital tools provide multimodality as semiotic mediation that are relevant in 21st century education. The notion multiliteracy entails the skill required to encode and decode language to read and write a language as a traditional point of view, while the notion digital multiliteracy invokes internet and technology-mediated literacy (see Elola & Oskoz, 2017 as the main review). Through digital multiliteracy, the interface among mode, modality, and media is inseparable in which mode as semiotic representation e.g., textual, aural, visual, is used to present information, while modality constitutes semiotic realization of one mode. Additionally, mode and modality might use asynchronous or synchronous computer mediated communication (CMC). The former might be conducted by writing in a *blog* which is read later by the reader for the discussion, while the latter could be carried out by using e.g., *Google Docs, Facebook* which entails a simultaneous chat between the writer and reader. Later, media constitutes technological means which combine visual (images), aural (sound), and textual (subtitle) modes (Guichon & Cohen, 2016).

Such digital tools provide new insight in English teaching and learning in the 21st century which creates new kinds of *genre* instead of classical ones e.g., description, argumentation, narration, exposition. The new genre is explicated by the use of texts via e.g., *wiki, blog, Facebook, Twitter* that enable language learners to use text, sound, and video simultaneously. Additionally, the use of video blogging (vlogging) enables learners to create

personal journals that later is shared via public websites such as YouTube, Facebook, or Twitter. This kind of new genre should be called for the additional elements of the curricular development at tertiary level. The emergence of this genre enables more elaborative and comprehensive discussion with reference to language making-meaning explicated in grammatical metaphor to make in depth-analysis on any kinds of semiotic mediation elucidated through digital multiliteracy.

In summary, GMs provide not only genre awareness (e.g., nominalization, verbalization, preposition) in terms of their stratal tensions between lexicogrammar and semantic discourse, but also genre disciplines (e.g., scientific texts, magazine, fictions) explicated as language register. Analyses on written and spoken languages across genre awareness and disciplines provide a comprehensive understanding of the language structure, use, and function. Likewise, the interface between task- and genre-based approaches might become powerful language pedagogy to implement at tertiary level. Eventually, incorporating ubiquitous information, communication, and technology create a new genre that needs to be discussed and elaborated for the future language research and teaching pedagogy in conjunction with GM use.

References

Byrnes, H. (2014). Task-based language learning: Insights from and for L2 writing: An introduction. In H. Byrnes & R. Manchon (Eds.), *Task based language learning: Insights from and for L2 writing* (pp. 1–26). Amsterdam, the Netherlands: John Benjamins.

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education, 20,* 50 – 66.

Byrnes, H., Crane, C., Maxim, H. H., & Sprang, K. A. (2006). Taking text to task: Issues and choices in curriculum construction. *ITL-International Journal of Applied Linguistics, 15 (2),* 85 –109.

Christie, F. & Derewianka, B. (2008). *School discourse: Learning to write across the year of schooling*. London: Continuum.

Christie, F. and Martin, J.R. (2007). *Knowledge Structure: Functional Linguistic and Sociological Perspectives*. London: Continuum.

Devrim, D.Y. (2015). *Teaching grammatical metaphor: Designing pedagogical interventions.* Newcastle upon Tyne: Cambridge Scholars Publishing

Elola, I. & Oskoz, A. (2017). Writing with 21st century social tools in the L2 classroom: New literacies, genres, and writing practices. *Journal of Second Language Writing, 36,* 52-60.

Fadilah, E. (2018). Rethinking the maintenance of CLT in Indonesia: A response to: Ariatna's (Vol. 7, No. 2, 2016) "The Need for Maintaining CLT in Indonesia". *TESOL Journal, 9 (1),* 224-236.

Guichon, N. & Cohen, C. (2016). Multimodality and CALL. In F. Farr, & L. Murray (Eds.). *The Routledge handbook of language learning and technology (*pp. 509-521). New York, NY: Routledge.

Halliday, M.A.K. (1985). *An introduction to functional grammar (1st ed.)*. London & New York: Edward Arnold.

Halliday M.A.K. (1994) *An introduction to functional grammar (2nd ed.)*. London & New York: Edward Arnold.

Halliday, M. A. K. (2004). *The language of science*: *Volume 5 (Collected works of M.A.K. Halliday)*. London, England: Continuum.

Halliday, M. A. K., & C. M. Matthiessen (2004). *An introduction to functional grammar* (3rd edn.). London: Arnold.

Halliday, M.A.K. (2009). Preface to Continuum Companion to Systemic Functional Linguistics. In M.A.K. Halliday & J.J. Webster (ed.), *Continuum companion for Systemic Functional Linguistics (pp.vii-viii)*. London: Continuum.

Halliday, M. A. K. & C. M. Matthiessen (2014). *An introduction to functional grammar* (4th edn.). London: Arnold.

Halliday, M.A.K. (1994). *Introduction to Functional Grammar* (2nd edition). London: Arnold.

He, Q., & Yang, B. (2018). A corpus-based study of the correlation between text technicality and ideational metaphor in English. *Lingua, 203, 51-65*

He, Q. & Wen, B. (2017) A Corpus-based Study of Textual Metaphor in English. *Australian Journal of Linguistics*, *37*(3), 265-285.

Hu, J. (2015). A comparative study of China English and Singapore English: the case of grammatical metaphor in academic discourse. *Journal of World Languages*, *2*(1), 50-61.

Liardét, C., L. (2013). An exploration of Chinese EFL learner's deployment of grammatical metaphor: learning to make academically valued meanings. *Journal of Second Language Writing*, *22*(2), 161–178.

Liardét, C. L. (2016). Grammatical metaphor: Distinguishing success. *Journal of English for Academic Purposes, 22*, 109–118.

Liardét, C. L. (2018). 'As we all know': Examining Chinese EFL learners' use of interpersonal grammatical metaphor in academic writing. *English for Specific Purpose, 50, 64-80.*

Martin, J.R. (2008). Incongruent and proud: de-vilifying 'nominalization'. *Discourse Society, 19,* 801-810.

Martin, J.R. (1992). *English text: System and structure.* Amsterdam: John Benjamins.

Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. *Journal of Second Language Writing*, 29, 51 –63

Ryshina-Pankova, M. & Byrnes, H. (2013). Writing as learning to know: Tracing knowledge construction in L2 German compositions. *Journal of Second Language Writing, 22 (2),* 179-197.

Ryshina-Pankova, M. (2010). Toward mastering the discourses of reasoning: Use of grammatical metaphor at advanced levels of foreign language acquisition. *Modern Language Journal*, 94, 181–197.

Taverniers, M. (2006). Grammatical metaphor and lexical metaphor: Different perspectives on semantic variation. *Neophilologus, 90,* 321-332.

Webster, J.J. (2009). An Introduction to Continuum Companion to Systemic Functional Linguistics. In M.A.K. Halliday & J.J. Webster (ed.), *Continuum companion for Systemic Functional Linguistics (pp.1-11)*. London: Continuum.

Yang, B. (2018a). Interpersonal metaphor revisited: Identification, categorization, and syndrome. *Social Semiotics*. doi: 10.1080/10350330.2018.1425322

Yang, B. (2018b). Textual Metaphor Revisited. *Australian Journal of Linguistics, 38*(2), 205-222.

Yasuda, S, (2017). Toward a framework for linking linguistic knowledge and writing expertise: Interplay between SFL-based genre pedagogy and task-based language teaching. *TESOL Quarterly, 51* (3), 576-606.

Yasuda, S. (2015). Exploring changes in FL writers' meaning-making choices in summary writing: A systemic functional approach. *Journal of Second Language Writing*, *27*, 105-121.

Yongsheng, Z. (2008). Nominalisation, verbalisation and grammatical metaphor. In J.J. Webster (Ed.), *Meaning in context: Strategies for implementing intelligent applications of language studies* (pp. 297–308). London: Continuum.

Video segments: A valuable tool for teaching English phonological processes

Bettiana Andrea Blázquez[*] - Universidad Nacional del Comahue, Argentina

ABSTRACT

This article explores the advantages of using authentic video segments - scenes, previews, documentaries, shorts and commercials- when teaching English phonological processes. The first section provides a possible classification of this type of audiovisual material and a description of its potential. Then, it offers a definition of phonological processes and a plausible typology. Finally, a framework for structuring a video lesson is outlined. In the second part two sets of sample activities, which have been designed on the basis of this framework, are shared and described. The overall aim of these worksheets is to view them as springboards for designing new activities in terms of other learning purposes. The employment of video segments will encourage learners to recognise phonological processes, develop their listening skill and extend their learning beyond the classroom.
Keywords: authentic video segments; phonological processes; listening; activities; EFL teaching

RESUMEN

Este artículo explora las ventajas del uso de segmentos de video auténticos - *escenas de películas, avances, documentales, cortometrajes* y *comerciales* - al momento de enseñar los procesos fonológicos del inglés. La sección inicial provee una posible clasificación de este tipo de material audiovisual y una descripción de su potencial. Luego, ofrece una definición de procesos fonológicos y una posible tipología. Finalmente, se esboza la estructuración de una clase de video. En la segunda parte, dos conjuntos de actividades, diseñadas en torno a este esquema de clase, se comparten y describen. El objetivo general de estas guías didácticas es el de visualizarlas como punto de partida para el diseño de nuevas actividades en función de otros propósitos pedagógicos. El empleo de segmentos de video alentará al estudiantado a reconocer procesos fonológicos, desarrollar la habilidad auditiva y extender el aprendizaje más allá del contexto áulico.
Palabras clave: segmentos de video; procesos fonológicos; habilidad auditiva; actividades; enseñanza de inglés como lengua extranjera

_____

* Corresponding author, e-mail: bettianablazquez@hotmail.com

THE PRESENT GLOBALISED world offers an array of video segments that may be advantageously incorporated into the English classroom as their combination of brevity, sound and images usually compels the learners' attention. The fact that this resource typically conveys contextualised messages makes it particularly valuable for teaching different features of natural speech such as rhythm, stress, intonation and modifications of sounds (Kelly, 2000; Rogerson-Revell, 2011; Underhill, 2005). Even though each of these features plays a leading role in the understanding of the English language, this last aspect, which is commonly referred to as "phonological processes" (Yoshida, 2016, p. 112) is the focus of attention of this article. These adjustments of speech result from native speakers' "articulatory short cuts when moving from one sound to the next" (Grant, 2014, p. 25) as undoubtedly, they do not speak word by word. To put it simply, words flow into each other and while "some of them maintain resemblance to the citation form, others are just pulled out of shape" (Cauldwell, 2012, p. 1). This variability between the expected pattern and the context-influenced form produced in connected speech makes learners face serious problems as listeners (Grant, 2014).

The impact of teaching connected speech processes on students' listening comprehension has been examined by Brown and Hilferty (1986, 2006), Ito (2006), Matsuzawa (2006) and Romanko (2008). They demonstrate that the instruction of these adjustments develops not only the students' perception of them, but also their overall listening skill. As regards English teachers' views on teaching phonological processes and other features of connected speech, a survey, conducted by Rogerson (2006), reveals that the majority of language teachers refrain from teaching these aspects. Their reasons range from not having appropriate materials to not knowing how to teach them. In the light of this evidence, she states the need for the development of more authentic teaching materials. In the same vein, Underwood and Wallace (2012), suggest exposing students to authentic sources of language in order to foster not only the learners' understanding of phonological forms, but also their confidence in the study beyond the learning setting.

The phenomenon described above puts in evidence the need for creating activities conducive to the recognition of these speech modifications. Having this purpose in mind, the aim of this article is to share two sets of classroom activities that have been designed on the basis of the exploration of video segments and phonological processes.

Theoretical Framework
Authentic Audiovisual Materials and Video Segments

Authentic audiovisual materials should basically be defined as "moving images, possibly accompanied by sound" (Taylor, 2009, p. 1) that have not been created for educational purposes. These filmed resources may be "animated or live action, scripted or spontaneous, factual or fictional, professional or amateur" (Taylor, 2009, p. 1). Far from threatening pronunciation teachers and learners, this new digital reality allows for the versatility and the

ease of access to audiovisual material, which was beyond reach years ago. Nowadays they are readily available on the Internet and on well-known free-distribution platforms such as YouTube or Google, which apart from offering full-length films, also share video segments like:

- Previews or trailers: movie-flashes that present condensed versions of future releases;
- Scenes: parts of films, drama series, soap operas or sitcoms;
- Documentaries: programmes about a particular topic which
- show facts and intend to inform the audience;
- Shorts: self-contained films that are approximately about 15 minutes in length;
- TV commercials: advertisements that are either short films or voice narratives of a product.

As an alternative source of input to traditional audio texts, pieces of filmed material have a number of points in favour. One of their most outstanding features is their brevity as their total running time is less than fifteen minutes, and thus they can be viewed several times in the pronunciation class. Besides, they are self-contained, i.e., they depict a situation that has a beginning and an end. This characteristic makes them function as freestanding texts, which leaves room for using them at any time in the lesson (Massi & Blázquez, 2008, 2010, 2012).

Video segments also exert compelling power, given that the eye is immediately caught and therefore curiosity about the meaning of words is aroused (Sherman, 2003). In addition, authenticity is an inducement, because "there is a special thrill in being able to understand and enjoy the real thing" (Sherman, 2003, p. 2). In this regard, Field (2008) points out that learners should have "the opportunity of hearing what L2 really sounds like" (p. 277). Furthermore, audiovisual texts prove to be a source of contextualised language that displays most of the characteristics of oral discourse, i.e., prosodic features, such as speed and volume, and paralinguistic features, including gestures, eye contact, proximity and the like (Massi & Blázquez, 2008).

According to Goodwin (2008) and Lowe (2007), the employment of clips in the pronunciation lesson allows learners to focus on segmental and suprasegmental features, mimic the multi-layered complexity of real-life interaction and develop listening comprehension skills. Videos enhance "the social significance of accents", because they are "the signature tune of a culture and should be heard in their social context" (Sherman, 2003, p. 53). This broad perspective reflects the contemporary use of English as an International Language (Jenkins, 2000; Walker, 2010) and goes in keeping with Marks and Bowen's (2012) line of reasoning about the importance of training the learners' "ears to expect variety" (p. 10). Thus, familiarity with different samples of the English language will eventually lead the learners to autonomy since they may apply what they have learnt in the classroom to other listening experiences in the outside world.

Phonological Processes

Understanding the English language involves a great load on processing the information that is being conveyed. One way of developing this ability is to raise learners' awareness of the different processes that operate in *connected speech*. The recent literature on the subject presents variation in what to include under this umbrella term, how to label its features and how to classify them. To take an example, Brown and Kondo-Brown (2006) and Brown (2012) use the term *reduced forms* or *reductions* to group, in general lines, contractions, strong and weak forms and phonological processes into connected speech. Another typical case is that some authors regard contractions as examples of the phonological process elision (Roach, 2009; Rogerson-Revell, 2011). Similarly, different terms are usually applied to make reference to the same type of phonological process. The use of blend, palatalisation, reciprocal assimilation and coalescence is a case in point. This brief account gives reasonable grounds for determining the scope of this article. To that end, firstly, it is necessary to define connected speech.

The term connected speech is "used in linguistics to refer to spoken language when analysed as a continuous sequence" (Crystal, 1980, p. 101). A similar definition is provided by Underhill (2005), who notes that "connected speech consists of a flow of sounds which are connected, grouped and modified" (p. 58). In sum, speech makes reference to a continuous string of phonemes "in which the ideal positions for the articulation of individual sounds may never be reached" (Dalton & Seidlhofer, 1994, p. 24) due to the influence of neighbouring segments that appear in context.

Native speakers, in an attempt to display their mastery of the language, economise their efforts in the articulation process, as the main objective is definitely to get their message across (Dalton & Seidlhofer, 1994). In doing so, they tend to: leave a sound out, insert a sound or/and make adjacent sounds more like each other (Dalton & Seidlhofer, 1994). Thus, the phenomena that imply eliminating, inserting or changing sounds in order to facilitate pronunciation are generally termed *phonological processes*. In effect, all languages distort sequences in connected speech; however, the way in which these processes operate is delimited by every language. Moreover, most native speakers are not familiar with these adjustments and, therefore, are often taken unaware when someone draws their attention to them (Collins & Mees, 2013). It should be noted that these modifications are phonological, because they affect sounds in context, i.e., within a word or at word boundaries (Dalton & Seidlhofer, 1994). Particularly, in this article, the connected speech phenomenon is restricted specifically to these adjustments of sounds that result from the application of phonological rules as in *take him* /teɪkhɪm/   /teɪkɪm/ rather than the memorization of lexical forms, for example, *will not   won't* /wəʊnt/ (Henrichsen, 1984). Accordingly, the phonological processes that typically occur in the spoken discourse of English and lie inside the scope of this article are elision, liaison or linking, assimilation and coalescence. These speech

modifications are detailed below, however, the list of categories provided in this article is not exhaustive:

- Elision takes place when a consonant or a vowel is omitted or left out. Some rules for elision are:
  - Omission of /h/: The glottal fricative is regularly elided from the weak forms of function words (Collins & Mees, 2013). This sound may be omitted in pronouns such as *he, him* or *her* as long as it does not appear in initial position in the tone unit. It may also be deleted from the word *have* when it functions as an auxiliary verb. For example: *He must have bought him a present* /himʌstəvbɔːtɪməprezənt/.
- Liaison or linking is the smooth connection of sounds. One of the rules for linking is:
  - Linking r: Some accents like Australian English or General British (GB) (Lindsey, 2017; Carley, Mees & Collins, 2017), are non-rhotic, i.e., speakers do not pronounce the /r/ after vowels, so they say /kɑː/, however they often use this phoneme to link words. This process requires modifications at the beginning and end of words in connected speech, as it consists in pronouncing word-final spelling r when the next word begins with a vowel. Common examples are: *four eggs* /fɔːregz/ and *later on* / ˈleɪtərɒn/. It is advisable to adopt this process, since it is an essential characteristic of fluent speech in GB (Finch & Ortiz Lira, 1982).
- Assimilation takes place when a given sound either takes on one phonetic feature of the following sound or affects this neighbouring sound (Celce-Murcia, Brinton & Goodwin, 2010). There are different types of assimilation in English, but, for practical reasons, only the most common type, assimilation of voice, is described:
  - Assimilation of voice occurs in "*has/have to* (when expressing obligation) and *used to* (when expressing former habitual action)" (Celce-Murcia, et al., 2010, p. 169). In these examples, the voiceless /t/ of *to* affects the voiced /z/, /v/ and /d/ and become voiceless /s/, /f/ and /t/ respectively. At the same time, in *used to*, the already assimilated voiceless /t/ causes the fricative voiced sound /z/ preceding it to assimilate and become /s/ (Celce-Murcia, et al., 2010). In *of course*, the voiceless /k/ of *course* conditions the preceding voiced fricative /v/ causing it to be pronounced /f/ (See Table 1 below).

Before the process	Assimilated form
Has to /hæz tə/	**/hæs tə/**
Have to /hæv tə/	**/hæf tə/**
Used to /juːzd tə/	**/juːs tə/**
Of course /əv kɔːs/	**/əf kɔːs/**

Table 1. Instances of regressive assimilation of voice (Adapted from Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. 2010).

The direction of this process may also be progressive as a given sound affects the voicing of the sound that follows. Examples of this type of assimilation in English are the regular plural, 3rd person singular and the regular past tense. In these cases the final sound conditions the voicing/voiced pronunciation of the ending (Celce-Murcia, et al., 2010) (See Table 2 below).

Conditioning sound	Assimilated form
Boards /d/	/bɔ:d➡z/
Talks /k/	/tɔ:k̩➡s/
Listened /n/	/ 'lɪsən̩➡d/

Table 2. Instances of progressive assimilation of voice

- Coalescence occurs when a first and a second sound in a sequence merge into a third sound with features from both original sounds (Celce-Murcia, et al., 2010). To take an example, in the assimilated form *don't you* / dəʊntʃu / the two postures, alveolar for /t/and palatal for /j/, coalesce into one palato-alveolar articulation /tʃ/. The coalesced /tʃ/ is widely heard within words too as in *actually* / 'æktʃuəli / and *Tuesday* / 'tʃu:zdi / (Lindsey, 2017). Table 3 below shows the possible contexts in which coalescence may take place.

Before the process	Phonological Context	Assimilated form
Don't you / dəʊnt ju /	/t / + /j/➡/tʃ/	/ dəʊntʃu /
Could you / kʊd ju /	/d / + /j/➡/dʒ/	/ kʊdʒu /
This year / ðɪs jɜ: /	/s / + /j/➡/ʃ/	/ ðɪʃɜ: /
These years / ði:z jɜ:z /	/z / + /j/➡/ʒ/	/ ði:ʒɜ:z /

Table 3. Instances of coalescence (Adapted from Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. 2010).

Framework and General Characteristics of the Sample Activities
The phonological processes mentioned above affect two fundamental aspects of oral communication: intelligibility and comprehensibility. The former refers to the extent to which a listener understands a speaker's message and the latter involves the amount of effort a listener makes to understand this message (Derwing & Munro, 2015). Certainly, the improvement of these speech dimensions usually becomes a desirable goal in any pronunciation classroom. Viewed this way, developing them partly entails a lot of training in the recognition of phonological processes (Celce-Murcia, et al., 2010; Jenkins, 2000; Gilbert, 2008). In

doing so, students generally engage in activities to accomplish a listening goal, during which they handle the realities of spontaneous speech (Cauldwell, 2013).

Several researchers have focused on the use of authentic audio and audiovisual material in improving listening comprehension in the language classroom. For example, Mendelsohn (1994), Mendelsohn & Rubin (1995) and Oxford (1990), among others, suggest incorporating natural audio material to prepare learners for real listening. Likewise, they demonstrate that learning occurs in three phases, namely, pre, while and post listening. Following this framework, Stempleski (2002) centers on the use of authentic audiovisual material in the language classroom and plans video-related lessons for three stages of activity: pre, while and post viewing.

The two sets of activities included in this article take video segments as a starting point. For this reason, these sequences follow Stempleski's (2002) approach in terms of the three phases mentioned earlier. This collection of activities are mainly aimed at integrating both meaning with form. In other words, rather than recognising phonological processes in isolation, the activities promote the identification of these forms in context. Another feature these activities have in common is that they are suitable for intermediate or upper-intermediate levels. The worksheets have been tailored for future teachers and translators who are attending the subject English Phonetics & Phonology I at Facultad de Lenguas, Universidad Nacional del Comahue, Argentina. As regards the accent variety that is used by the speakers, both videos show instances of GB. Yet, our learners are progressively exposed to other varieties of English, such as Standard American, Australian and Scottish accents, to name a few, with the purpose of developing our learners' receptive skills. As far as previous knowledge is concerned, it is important to consider that these sets of activities imply the mastery of the International Phonetic Alphabet (IPA) and the rules of phonological processes. The worksheets have been designed to give the learners practice in identifying phonological processes for them to develop their listening skill. However, they may function as templates for adaptation in terms of other pedagogical purposes.

Description of Sample Activities

The first set of activities (see Appendix A) is based on a scene taken from the film *Love, Rosie* (2015). The criteria for selecting this clip entailed a careful examination of the exploitability of the material, i.e., its content, its linguistic level, its visual support, its speech delivery and the accent used by the speakers. The first True/False activity gives the students the opportunity to deal with content. It belongs to the *pre-viewing stage* as it implies some preliminary work, during which the learners activate prior knowledge and make predictions about the video segment they are going to watch. This pre-viewing phase also encompasses a set of warming-up questions about the content of the scene and the theoretical point that is being consolidated, i.e., phonological processes. In this way, the learners' attention is drawn to this particular feature of the target language so as to allow the students to approach the

video with certain schemata later on. Both activities create an atmosphere for watching and guide the learners to view the scene with a purpose in mind.

The first activity of the *while-viewing stage* involves watching the preview of the film. This condensed version provides the learners with visual input concerning, among other aspects, paralinguistic features such as body movements, gestures and clothes people wear. The visual information as well as the speakers' interaction and background music help the students to grasp meaning and check what they have anticipated. Having discussed the content of the film, confirmed or rejected their hypotheses about the reasons why the couple is having an argument and reflected on the features of connected speech and phonological processes, the learners are ready to take the following step. The second activity invites them to focus on specific information as the students are asked to pay attention to some words or phrases in bold and recognise phonological processes. Specifically, the learners are prompted to record this information by transcribing the speakers' pronunciation of the selected words or phrases. The purpose of this phase is, then, to guide the learners into the viewing process so as to foster perception and comprehension (Massi & Blázquez, 2008, 2010, 2012).

As to the *post-viewing phase,* the first activity invites the learners to classify phonological processes by establishing connections between their previous knowledge and the new information. Additionally, this task attempts to encourage the students to analyse what they have seen or heard in the light of the knowledge of and about the phonological aspect that is being consolidated. Finally, an extending activity is suggested for students to transfer key forms of the video segment to an analogous situation, in this case a letter. In this activity, the learners have access to the audiovisual material of the letter being read by one of the main characters of the film. In this way, the learners are expected to form associations between meaning and other aspects, namely phonological rules, spelling and pronunciation.

The second set of activities (see Appendix B) centers on the preview of the film *Cinderella*. This audiovisual material was chosen on the basis of the students' familiarity with the fairy-tale, complexity of structures and vocabulary, clarity of speech, quality of sound and picture, speed and the like. In the first multiple-choice activity the learners concentrate on meaning and speculate about the characters, their actions and the setting. After that, the learners are involved in a True/False pair work activity in which they are required to pay attention to the connected speech phenomenon that is being learnt. Both activities comprise the *pre-viewing stage* as the students get ready to gain information from visual and/or auditory clues and associate it with their previous knowledge. The main objective of this phase is to make the most of learners' predictive skills by laying the ground for creating expectations and generating a reason for viewing the clip (Massi & Blázquez, 2008, 2010, 2012).

During the *while-viewing stage* the learners are invited to watch the preview so as to check their predictions as well as select part of the incoming information. Specifically, the

students are engaged in a gap-fill activity. The words or phrases that have been extracted from the text have something in common: all of them display phonological processes. This type of activity works as a scaffold to the recognition of this aspect of pronunciation and the comprehension of the language used in this condensed version of the film.

In the *post-viewing section* the learners are asked to transcribe the chunks they perceived in the previous stage and detail the kind of phonological processes these words or phrases depict. In doing so, the students use their prior knowledge as well as the new information to establish connections. It is important to emphasise that this activity is intended to embark the learners on noticing the salient features of the target language by consciousness-raising. The final consolidating task implies some extra cognitive effort as the learners are encouraged to view an interview with Cate Blanchett and jot down notes on her views on taking up part of this film without receiving any orthographic help. The emphasis is first on understanding real language, and then on form. This outcome is further built on at a later stage by inviting the students to discriminate between two possible ways of pronouncing some tone units, i.e., with or without a specific phonological process, and encouraging the learners to find more instances of modifications of speech. The objective of this final step is twofold: the learners not only improve their perception, but also foster the connection between the discovery of phonological contexts and the rules underlying the linguistic input the learners handle.

Conclusion

Mastering individual sounds or segmentals is not sufficient to learn the pronunciation of a language. According to Carley, Collins and Mees (2017), there is more to that, as students should also pay attention to other aspects like assimilation, liaison and elision, among others. Video segments display these adjustments of connected speech and prove to be a motivating alternative in the pronunciation lesson. Moreover, their brevity allows for repeated viewings, which are fundamental to identify phonological processes, develop listening comprehension as well as fully acquire other aspects of the target language.

The sets of activities presented in this article endeavor to provide possible ways of exploiting this audiovisual resource when teaching phonological processes. They also build up the learners' "right expectations about the kind of sound patterns they are likely to be confronted with in normal native speech" (Dalton & Seidlhofer, 1994, p. 116). It goes without saying that these activities by no means exhaust the various possibilities of employing video segments in the pronunciation and EFL lesson. Exploiting the wide diversity of the short videos will eventually guide learners into recognising phonological processes, developing their listening skill and gaining autonomy. The ultimate goal is to help students experience intelligibility and comprehensibility when listening to authentic English speech.

References

Brown, J. D. (2012). *New ways of teaching connected speech*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Brown, J. D., & Kondo-Brown, K. (2006). Introducing connected speech. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 3-15). Honolulu: University of Hawai'i Press.

Brown, J. D. & Hilferty, A. (1986). Listening for reduced forms. *TESOL Quarterly*, 20(4), 759-763.

Brown, J. D., & Hilferty, A. (2006). The effectiveness of teaching reduced forms for listening comprehension. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 51-58). Honolulu: University of Hawai'i Press.

Cauldwell, R. (2012). *The two-sides rule in teaching listening and pronunciation*. Retrieved September 10, 2018, from

http://www.developingteachers.com/articles_tchtraining/two_sides1_richard.htm

Cauldwell, R. (2013). *Phonology for listening*. Birmingham: Speechinaction.

Carley, P., Collins, B., & Mees, I. M. (2017). *English phonetics & pronunciation practice*. Oxon: Routledge.

Celce-Murcia, M., Brinton, D. M. & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide*. Cambridge: Cambridge University Press.

Collins, B. & Mees, I. M. (2013). *Practical phonetics & phonology*. Oxon: Routledge.

Crystal, D. (2008). *A dictionary of linguistics and phonetics*. (6th ed.). Oxford: Blackwell Publishing.

Dalton, C. & Seidlhofer, B. (1994). *Pronunciation*. Oxford: Oxford University Press.

Derwing, T. & Munro, M. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia: John Benjamins Publishing Company.

Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Finch, D. & Ortiz Lira, H. (1982). *A course in English phonetics for Spanish speakers*. London: Heinemann.

Gilbert, J. B. (2008). *Teaching pronunciation. Using the prosody pyramid*. New York: Cambridge University Press.

Grant, L. (2014). *Pronunciation myths*. Michigan: University of Michigan.

Goodwin, J. (2008). *Using film clips to teach pronunciation*. Retrieved September 10, 2018, from http://joelswagman.blogspot.com/2015/06/using-tv-and-movie-clips-to-teach.html

Henrichsen, L. E. (1984). Sandhi variation: A filter of input for learners of ESL. *Language Learning, 34*, 103-126.

Ito, Y. (2006). The comprehension of English reduced forms by second language learners and its effect on input-intake process. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 67-84). Honolulu: University of Hawai'i Press.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Kelly, G. (2000). *How to teach pronunciation*. Harlow, England: Pearson Education Limited.

Lindsey, G. A. (2017). *English after RP*. London: Tequisté TXT.

Lowe, M. (2007). Films in English Language Teaching. *IH Journal*, *23*, 16-19.

Marks, J. & Bowen, T. (2012). *The book of pronunciation. Proposals for a practical pedagogy*. Peaslake: Delta Publishing.

Massi, M. P. & Blázquez, B. A. (2008). Exploiting DVDs' Extra Features: An Added Bonus in the EFL Class. Humanising Language Teaching (HLT). Free Online EFL Magazine for Teachers, 10 (4). Retrieved September 10, 2018, from
http://old.hltmag.co.uk/aug08/mart.htm

Massi, M. P. & Blázquez, B. A. (2010). The Shorter the Better: Using Shorts in ELT. Modern English Teacher, 19 (2), 36-42.

Massi, M. P. & Blázquez, B. A. (2012). A Short is Worth a Thousand Films! Teaching English with technology (TEwT). Free Online Journal for TESOL Teachers, 12(3), 62-83. Retrieved September 10, 2018, from https://www.ceeol.com/search/article-detail?id=148649

Matsuzawa, T. (2006). Comprehension of English reduced forms by Japanese business people and the effectiveness of instruction. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 59-66). Honolulu: University of Hawai'i Press.

Mendelsohn, D. J. (1994). *Learning to listen: A strategy-based approach for the second-language learner*. San Diego: Dominie Press.

Mendelsohn, D. J. & Rubin J. (1995). *A guide for the teaching of second language listening*. Singapore: Dominie Press, Inc.

Oxford, R. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House.

Roach, P. (2009). *English phonetics and phonology. A practical course*. Cambridge, United Kingdom: Cambridge University Press.

Rogerson, M. (2006). Don'cha Know? A survey of ESL teachers' perspectives on reduced forms instruction. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 85-98). Honolulu: University of Hawai'i Press.

Rogerson-Revell, P. (2011). *English phonology and pronunciation teaching*. London, Great Britain: Bloomsbury Publishing.

Romanko, R. (2008). Whaddaya Think about teaching reduced forms? *Presentation at the JALT 2008 Annual Conference*, Tokio.

Sherman, J. (2003). *Using authentic video in the language classroom*. Cambridge: Cambridge University Press.

Stempleski, S. (2002). Video in the ELT classroom: The role of the teacher. In J. C. Richards & W. A. Renandya, *Methodology in language teaching: An anthology of current practice* (pp 354-367). Cambridge: Cambridge University Press.

Underhill, A. (2005). *Sound foundations: Learning and teaching pronunciation* (2nd ed.). Macmillan Publishers Limited.

Underwood, P., & Wallace, M. (2012). The effects of instruction in reduced forms on the performance of low-proficiency EFL university students. *The Asian EFL Journal*, *14*(4), 1–24.

Taylor, S. (2009). Using Video: Theory. *IH Journal*. Issue 26. Spring. Retrieved September 10, 2018, from http://ihjournal.com/using-video-theory

Walker, R. (2010). *Teaching the pronunciation of English as a lingua franca*. Oxford: Oxford University Press.

Yoshida, M. T. (2016). *Beyond repeat after me: Teaching pronunciation to English learners.* Alexandria, USA: Tesol Press.

Appendix A

<div align="center">

LOVE, ROSIE

</div>

Type of Video Segment: Scene
Runtime: 3 minutes
Level of English: Intermediate to Upper-Intermediate – Background knowledge of phonological processes
Online resources: https://www.youtube.com/watch?v=SqSE6Kzuht0
https://www.youtube.com/watch?v=6bpgK9fXGkQ
https://www.youtube.com/watch?v=DoKPEWa4n5A

PRE-VIEWING

1. **Focus on Meaning**
   Are the following statements True (T) or False (F)?
   a. Rosie and Alex are girlfriend and boyfriend.
   b. They both live in England and plan to move to Boston, North America.
   c. Alex is the only one who flies to North America to go to college.
   d. She is pregnant and Alex is the baby's father.

2. **Focus on Form**
   In the scene you are going to watch Rosie and Alex are having an argument.
   Answer the questions below:
   a. Why do they quarrel?
   b. What's their speech like? Do they speak word by word?
   c. What phonological processes are Rosie and Alex likely to use?

WHILE-VIEWING

1. **View the preview of *Love, Rosie*** https://www.youtube.com/watch?v=SqSE6Kzuht0 **to check your predictions.**

2. **Watch a scene from the film *Love, Rosie*** https://www.youtube.com/watch?v=6bpgK9fXGkQ
   Check your predictions and pay special attention to the words or phrases in bold.
   What phonological processes do Alex and Rosie produce? Complete the boxes with the corresponding pronunciation. Use phonetic script.

Alex: Rosie! Rosie, where are you going?

Rosie: Back home. And I mean England. This is too crazy for me, okay? I walked straight into the middle of something here!

Alex: Look, **it's** all fine, honestly. Sally and I just needed to air some stuff!

Rosie: Oh, good, I'm **glad you're** fine, because I'm not! We were out together the whole night, Alex! And all the while your girlfriend was at home expecting a baby! **And you** didn't say!

Alex: Come on, Rosie. Nothing **happened**, we were just...We were just having some fun.

Rosie: **That's** what you got me over here for then? Just some fun?

Alex: I wanted to see you! For Christ sake, you're my best friend!

Rosie: Or maybe you needed someone from **your old** life to point out the truth!

Alex: What what truth?

Rosie: **You're in** a mess, Alex! This whole **situation** is one big bloody giant mess!

Alex: Me? I'm in a mess?

Rosie: And **you're in** denial!

Alex: You're projecting!

Rosie: Okay, can we just stop this psychobabble here and just talk like English people, please!

Alex: Fine! Fine! Fine. Fine. Look. I was worried that you might find it hard, you know, coming here.

Rosie: Hard?

Alex: I mean, you've had a tough time, Rosie, and seeing where we live, you know, our apartment, our lifestyle, **it's** just...

Rosie: Okay! You thought I was jealous of you? No. I pity you, Alex! All I see is someone compensating **for a** crap personal life by schmoozing **his** way up the career ladder!

Alex: Crap personal life? Ah **let's** see. Um, stable relationship with beautiful girlfriend, lots in common, a great circle of **friends**, a kid on the way, **who's** gonna have two **parents**, by the way, not one! Rosie, come on, I...

Rosie: Just forget it, Alex! Go ahead and climb your greasy pole!

Alex: You know, I will!

Rosie: Just don't come whinging to me when you realise **you're empty** inside.

Alex: No **danger of** that!

Rosie: Good!

Alex: Excellent!

Rosie: Fantastic!

POST-VIEWING

1. **Place the different words and phrases in the bubbles below.**

ELISION
....
....
....
....

LINKING
....
....
....
....

COALESCENCE
....
....
....
....

ASSIMILATION
....
....
....
....

2. **At home follow this site** https://www.youtube.com/watch?v=DoKPEWa4n5A
   Listen to Rosie reading a letter sent by Alex a long time ago. Answer the questions and do the activity below.
   a. What is Alex trying to say to Rosie?
   b. What phonological processes does Rosie use while reading this letter?
   c. **Transcribe the whole letter in phonetic script. Then, record yourself reading it and upload the file to the Moodle platform.**

Rosie,

   You deserve someone who loves you with every beat of his heart. Someone who will always be there for you, and who will love every part of you, especially your flaws. I know Katie needs her dad. I don't want to intrude. So if this is all wrong, just ignore it and I promise I'll never broach the subject again. Greg's not the man for you, Rosie. Twice I've let you slip through my fingers. Let's stop being afraid and take the chance. I know now I can make you happy. Call me if you feel the same way.

Love, Alex

Appendix B

<div align="center">

CINDERELLA

</div>

Type of Video Segment: Preview
Runtime: 2. 20 minutes
Level of English: Intermediate to Upper Intermediate – Background knowledge of phonological processes
Online resources: **https://www.youtube.com/watch?v=20DF6U1HcGQ**
https://www.youtube.com/watch?v=WVZQUqGSBw4

PRE-VIEWING

1. **Focus on Meaning**
   Circle the corresponding option.
   a. Cinderella is about a girl named Ella[1]/Bella[2] who finds herself at the mercy of her cruel stepmother and stepsisters.
   b. She follows the upbeat mantra that her loving mother passed along to her: "Have faith[1]/courage[2] and be kind."
   c. The girl's fortunes begin to change when she meets the prince[1]/the mice[2].

2. **Focus on Form**
   Work in pairs. Are the following statements True (T) or False (F)?
   a. Phonological processes take place when native speakers articulate with maximal efficiency rather than clarity.
   b. Phonological processes can only occur at word boundaries.
   c. The different types of phonological processes in English are elision, liaison or linking and coalescence.

WHILE-VIEWING

Watch the preview of Cinderella https://www.youtube.com/watch?v=20DF6U1HcGQ
to check your predictions. Then, view it again and fill in the gaps.

MOTHER: I want to tell you a secret that will see you through all the [1]……………… that life can offer. Have courage and be kind.
FATHER: She'll merely be your stepmother and you´ll have two lovely [2]……………… to keep you in company. So I'll know as [3]……………… [4]……………… I may be [5]……………… [6]……………… be safe.
STEPMOTHER: [7]…………………………… [8]…………………… prefer to eat when all the work is done, Ella?

CINDERELLA: Yes, stepmother.

STEPMOTHER: You needn't call me that! Madam will do.

STEPSISTERS: Cindy wench, dirty Ella, Cinderella.

PRINCE: Are you all right? Miss, what do they call you?

CINDERELLA: Never mind what they call me.

PRINCE: I'm sorry.

CINDERELLA: [9]………….………… not your doing.

PRINCE: Nor yours either, I'll bet. I [10]………..……… [11]…….……… see you again, miss.

CINDERELLA: And I you.

MOTHER: Ella you have more kindness in your little finger than most people possess in their whole body.

CINDERELLA: It was my [12]……………….…………. old dress.

STEPMOTHER: It would be an insult to take you to the palace [13]……………………… in these old [14]…………….……………. You shall not go to the ball.

FAIRY GODMOTHER: Excuse me why are you crying?

CINDERELLA: Who are you?

FAIRY GODMOTHER: [15]………..………… better.

CINDERELLA: My fairy godmother.

FAIRY GODMOTHER: Now, where's mice?

CINDERELLA: They're made of glass?

FAIRY GODMOTHER: And you'll find they're really comfortable. Now off you go for you shall go the ball.

CINDERELLA: They [16]……….…… [17]……….…… looking [18]………. [19]……………….

PRINCE: Believe me, they [20]……….…… [21]………… looking [22]………. [23]……….……

MOTHER: Where [24]………..…… [25]……….…… kindness, [26]………..……… [27]……….…… goodness. And where [28]………..….…… [29]……..……… goodness, [30]………..…….……… [31]……………..… magic.

PRINCE: I [32]………..……. [33]……..……… see [34]………..……… [35]………………….…!

STEPMOTHER: Are you looking for this.

POST-VIEWING

1. **Watch the clip again and pay special attention to the chunks you have completed. In the chart below transcribe these phrases and identify phonological processes.**

TRANSCRIBED PHRASES	PHONOLOGICAL PROCESSES

1. **As homework, browse this site** https://www.youtube.com/watch?v=WVZQUqGSBw4 **Watch an interview in which Cate Blanchett, the actress who performed Cinderella's stepmother, answers a question about her attraction to the project and her love for fairy-tales.**

   a. **Reconstruct her opinion. Jot down notes on why she loves fairy-tales, why the world can be a nasty place and why she was excited to be part of the telling of this story.**

………
………
………
………
………
………
………
………

b. **Circle the correct answer. Does Cate Blanchett say…**

1. /aɪ stɪl riːd ðəm /ˌsɪndəˈrelə r ən ˈbjuːti ən ðə biːst/	**2.** /aɪ stɪl riːd ðəm/ˌsɪndəˈrelə ən ˈbjuːti ən ðə biːst/
1. /aɪ lʌv ˈfeərɪteɪlz/ ən ˌsɪndəˈrelə r ɪn pə ˈtɪkjʊlə/	**2.** /aɪ lʌv ˈfeərɪteɪlz/ ən ˌsɪndəˈrelə ɪn pə ˈtɪkjʊlə/
1. /ðə greɪt θɪŋ əˈbaʊt ju nəʊ/	**2.** /ðə greɪt θɪŋ əˈbaʊtʃu nəʊ/

c. **Now find more examples of phonological processes and share them with a partner.**

EAP in multi-discipline classes: An experience

Florencia Beltramino[*] - Universidad Católica Argentina, Argentina

ABSTRACT

This article will share how a non-curriculum-planned English for Academic Purposes (EAP) approach was designed, sneaked in, instrumented and assessed in an English for General Purposes (EGP) undergraduate course in a multi-discipline class at undergraduate level in Argentina.

Keywords: EAP; multi-discipline; motivation; materials

RESUMEN

Este artículo comparte la manera en que se diseñó un enfoque no planificado para inglés con fines académicos, y cómo fue incluido, implementado, y evaluado en un curso universitario de inglés con fines generales en una clase multi-disciplinaria en Argentina.

Palabras clave: EAP; multi-disciplina; motivación; materiales

* Corresponding author, e-mail: florenciabeltraminoprof@hotmail.com

AS TEACHERS AT educational institutions, we are not always in a position to make all the decisions regarding the classes we conduct due to the fact that complex multifaceted institutional aspects penetrate curriculum design and impact on its implementation. There are moments, however, when teachers *feel* that trying an alternative approach may prove to be a suitable course of action in order to move from a static curriculum conceived as a finished product to a dynamic and constantly-evolving one conceived as a process (Knight, 2001) with the ultimate objective of better adapting to students' needs. This article will share how a non-curriculum-planned English for Academic Purposes -EAP-approach was designed, sneaked in, instrumented and assessed in an English for General Purposes -EGP- undergraduate course in a multi-discipline class at undergraduate level in Argentina.

The Context

English is a mandatory annual subject for all degrees at the Universidad Católica Argentina (UCA), Facultad Teresa de Ávila, with two levels, name it English I and English II by the end of which undergraduates achieve an A1 English level according to the Common European Framework of Reference (CEFR) for Languages. Usually students take English I between their first and second year at the university and English II between their third and fourth year of studies. Initially, English courses at the UCA were discipline-oriented so that Law students would attend a Legal English reading-comprehension class, Psychology and Educational Psychology students would have their purposefully-designed reading-comprehension English course and so would do International Affairs and Political Science students. In time, the University entered an expansionist process by means of which four more degrees were incorporated to its educational offer (Notary Public, Accounting, Economics and Human Resources Management) resulting in a significant increase in students enrollment and a growing complexity in fields of study. At present, the number of students in an English class can be as high as 60 people at the beginning of the academic year though an average of 40 reach the second semester. There is only one EFL tenured teacher per course, however, TWO graduate teacher assistants (GTA) had been incorporated the year this project was implemented, 2017. Facing this new context, the University authorities instructed that the foreign language teaching approach should be EGP so that in one single English class students of different disciplines could take the same course. The underlying reason was that schedule arrangements for offering discipline-oriented courses for all new fields of study could not be possibly implemented.

Students' Needs

English is the chosen foreign language at the UCA as well as most universities from non-English-speaking countries as part of their curricula for simple and powerful reasons: it is the language of science and the modern world's *lingua franca*. At long last, its teaching at higher education should provide students with a handful of tools for accessing,

understanding and processing academic and/or professional contents of their interest and necessity, and, ideally, for producing knowledge if an international scope is pursued. Considering that English teaching at the UCA was covering general purposes but not specific ones the EFL practitioners understood that undergraduates might be lacking some skills which allow for academic or professional genre understanding in the foreign language. Analogously, a question raised: Did students consider that being able to comprehend and process academic pieces in English was an asset to acquire at University? So as to give a proper answer a survey was conducted whose sample covered 90% of the students' population attending English I and English II. Most of them, 53.9%, expressed that they would like to have a blended EGP and EAP instruction; 29.7% would prefer just EGP, and 16.4% considered that just EAP was the desirable choice. Given the facts, teacher's and students' perception intersected in that EAP instruction was needed to complement the current EGP approach.

The Procedure

Embedding an EAP approach to complement a multi-disciplinary course which was originally designed as an A1 general-purpose class posits considerable challenges to EFL lecturers: discipline specificity, language level, content selection and time constraints.

Discipline Specificity

EAP is agreed to have certain characteristics such as being goal oriented and subject specific. It assumes some previous knowledge of the L2 and is designed to meet learners needs in relation to their academic or professional environment (Robinson, 1991). One of the major feats is detecting what specific contents undergraduates of different fields of study need to handle and how to bring them together into a single classroom setting.

Subject specificity was tackled following what Dudley-Evans and St John (1998) suggested is the common-core of EAP, i.e. cross curricular subjects, text genres and text functions which serve as the basis for groups of disciplines. To exemplify, Law, International Affair, Notary Public and Political Sciences have much in common. A similar scenario occurs with Psychology and Educational Psychology, or among Accounting, Economics and Human Resources Management. Thus, the scope can be reduced to three broad branches instead of nine specific ones, making it possible to address the English teaching approach needed for academic or professional purposes.

Students' L2 level

Despite the fact that 95% of the undergraduates the UCA had the subject English across all the years of their secondary studies, very few can achieve a level higher than A1. At the beginning of their studies undergraduates are offered the possibility of taking a computer-based English achievement test. It mainly consists of multiple-choice exercises that assess

communication skills, grammar, vocabulary and reading comprehension corresponding to an A1 plus level. It is a little above a regular A1 level since it incorporates some, though not most, features of an A2 level. Passing this test, which can be taken just once during the degree, means being freed from attending English I and English II and getting a passing mark. Most novices decide not to take it, however, maybe because they are not confident enough to even trying it. It must be noticed that failing to achieve a passing mark in this exam does not bring about any consequences for students because its results are neither publicly displayed nor registered in academic records. From those who do take the test, only a 35% passes it. Consequently, undergraduates' English entry level could be named as a starters level, though certainly their interlanguages differ from person to person.

For the sake of embedding EAP into the EGP course it seemed sensible to do it towards the end of the English II course so that all students would have at least a basic understanding of the L2, which facilitates grasping more complex language chunks such as the ones present in academic or professional contexts.

Content Selection and Time Constrains

Around reaching the end of the second English course and having completed the syllabus some time before the expected date, the remaining four classes of the academic year were disposed for trying an EAP teaching intervention.

Having only a few classes available thoughtful content planning was mandatory. An EAP experienced lecturer and author was consulted, N. A. Gigena of Universidad Nacional de Entre Ríos. In terms of texts selection Gigena (personal communication, October 17, 2017) coincided with Dudley-Evans and St. John in that common-core texts are preferable in the described teaching-learning context. Additionally, the scholar suggested resourcing from texts of fields such as Anthropology, Philosophy or Ethics, which are cross-curricular, to introduce key features which she regards as essential to academic/professional genres comprehension:

- noun phrase and verb phrase analysis and resolution in Spanish
- modal verbs appraisal, and
- -ing cases correct interpretation and translation into the L1

According to Gigena, learners' motivation is the driving force behind the teaching approach and contextualization of EAP courses. As she stated (Gigena, 2010, p. 431),

> [students' motivation] is the reason why EFL practitioners should efficiently integrate into the L2 curriculum objectives and tasks which can be perceived as valuable by learners. This will be achieved if the objectives and tasks succeed in reflecting students' needs and interests in connection with the degree they pursue.

Implementation

Initially, an authentic descriptive text about Anthropology was selected and reading comprehension activities were designed as an introduction to the present the new contents. The same text was used for all three broad branches which will be referred as Law, Psychology and Economics. The exercises students had to resolve were a traditional True, False of Not in the Text activity followed by comprehension questions produced in the L2 but which were meant to be answered in the L1. Especially this last type of task presents a significant challenge: interpreting in the L2 and translating into the L1.

Even when able to understand the general idea of the text, giving more precise answers in Spanish proved to be hard for the learners. This difficulty was the trigger to introduce the planned contents and some translation techniques. First, the noun phrase was introduced, considering cases of pre and post-modification and its resolution in growing levels of complexity. A simple chunk was extracted from the text which served as a kick start, then gaps were given for students to complete which implied applying their previous knowledge and new items, the ones to be introduced, were already given. This can be better understood with an extract from the exercise whose aim was teaching noun phrase pre-modification translation:

1. human societies (line 3): sociedades _____
2. modern human societies: sociedades _____ _____
3. these modern human societies: _____ sociedades _____ _____
4. those old and modern human societies: _____ sociedades _____ _____y antiguas
5. studied human societies: sociedades _____ _____
6. extensively studied human societies: sociedades _____ ampliamente _____
   sociedades _____ _____ ampliamente
7. developing countries human societies: las sociedades _____ de _____ en desarrollo

In number 1 above the phrase *human societies* was retrieved from the text as indicates the parenthesis which shows its line location. Students were told that to start translating they had to identify the head of the phrase which could be detected by answering to the question: What is the phrase *talking* about? The head was then underlined by students so as to remember it was the starting point. The next sentences though not directly from the text, are adaptations of the original version which were designed to present aspects of a noun phrase pre-modification that can be challenging such as word order in examples 2, 3 and 4, the past participle *studied* in number 5, the possibility of the double location of the adverb *extensively* in number 6, and the translation of the present participle *developing* in phrase number 7. However, elements of the phrase which students were already acquainted with were left for them to complete, for example the demonstratives *these* and *those*. Once a new

element was introduced, it was then not given resolved in the next phrases, for example the past participle *studied* was given in number 5, but not in its next appearance in number 6. The rationale for such a methodology lies in Vygotsky's zone of proximal development theory (as cited in Daniels, 2005) which recommends departing from prior knowledge to increasingly incorporate new contents in growing degrees of complexity.

The same methodology was implemented for the presentation of techniques to identify, understand and translate noun phrases post-modification, simple and complex verb phrases including passive voice, and modal verbs. The work with these contents occupied two of the four classes available.

For the third class students were divided into the three broad fields Law, Psychology and Economy to be assigned authentic descriptive texts according to each of the mentioned disciplines, yet the type and number of tasks was the same: answering comprehension questions in the L1, a true or false, then resolution of noun phrases and verb phrases followed by the translation of a paragraph. Undergraduates worked in peers, and once finished, they were urged to contrast their answers with other students' to build their self-confidence. After that, each member of the teaching team, i.e. the tenured teacher plus the 2 GTAs, took charge of a subject field to check the answers by alternating between asking for volunteers and randomly choosing learners to answer. At the end of the class the learners were asked to gather in groups of four or five and select a book review in English connected to their professional interest and e-mail it to the professor who assessed their suitability for working with them the next and last lecture. Book reviews seemed the natural next step to descriptive texts due to being one of the shortest academic genres. Since elaborating activities for several different texts - there was a total of 9 - is often a hard-to-accomplish challenge due to time-management reasons, generic questions were produced that applied to all book reviews alike. All questions were written in English but asked to be responded in Spanish. Some examples were:

Who wrote the book?
What is the name of the book?
Who is the reviewer?
What is the book about?

Next, students were instructed to translate into Spanish the topic sentence of each paragraph. To conclude, each group had to express whether the reviewer had a positive impression of the book and whether they would find it useful for their studies and consider buying it. As a follow-up, learners socialised the tasks resolutions with the whole class making use of ICTs. Power Point or Prezi presentations were elaborated either in the classroom with students' own mobiles or at the university's computer lab.

The assessment of such a variety of texts being all instrumented in parallel just in a single class resulted to be a huge and exciting challenge at the same time. In the lapse of two

weeks -which was the time span between class 3 and class 4- each teacher read 3 out of the 9 reviews and produced suggested keys. That process involved going deep into unfamiliar specialist terms and concepts to be able to give proper responses to students' possible doubts. During the fourth class, learners worked on the texts in teams and after elaborating the ICT-mediated answers shared them first with the teacher who had dealt with their topic and next with the whole class.

Conclusion

It is too well known that teaching-learning contexts are anything but simple scenarios. Moreover, when the content to be taught and learnt is a language, both protagonists -learners and teachers- are presented with an extra challenge, that of the code of communication. Notwithstanding class size, all English courses entail a certain level of disparity between students' language goals and attainment levels. However, as Pulverness acknowledges it, the "[d]emands on the teacher common to any teaching situation -in terms of classroom management, involving and motivating students, attending to individual learning styles- are magnified exponentially in the large class" (2017, p. 2). While attempting to cater for the English linguistic needs of undergraduates in large mixed-discipline courses, as it is the case at UCA, Pulverness' assertion becomes premonitory.

What triggered the search for an alternative, meddled EFL teaching approach was the perception that undergraduates needed to be given resources that would empower them to successfully cope with career-specific materials. It goes without saying that being able to read authentic texts is not enough in covering all future graduates' needs. It was, however, the path chosen according to content prioritization and time constraints in an attempt to bridge the gap between the official curricula and uncovered contents which the EFL lecturers perceived as crucial for future professionals.

Since there is an unresolved dispute about EAP teaching methodology (Pulverness, 2017), it seemed sensible to implement some guidelines proposed by Dudley-Evans and St John's common-core approach and Gigena's valuable experience in contexts closer to UCA's reality. Thus, text genre selection reflected typical job-related pieces in English and content presentation was triggered by the need of achieving reading comprehension. The fields of study were grouped into three broad categories with shared groundings to obtain a nearer proximity to subject specificity. Students' engagement was encouraged by making them select texts of their own interest and assess their purpose, understanding that "given the difficulty of identifying the right material the best source of such texts, may well be the students themselves" (Pulverness, 2017, p. 7). Technology was an ever present aid across the whole process. The use of devices such as e-dictionaries, translators, Power Point and Prezi software among others, was not only allowed but fostered. We understand that the ICTs must take part in the EFL class for two main reasons. On the one hand, students will profit from

their usefulness; on the other, they will become aware that tech tools cannot be blindly relied upon but demand users' constant and careful supervision.

To sum up, it can be said that implementing an EAP approach in a mixed-discipline undergraduate class with an original EGP curriculum demanded extra time and dedication in terms of teaching and assessment strategies, material selection and teacher training, yet it was a most rewarding experience. After classes had finished students answered a survey about the relevance they assigned to the contents studied throughout the whole academic year in the English course. More than half of them, a 56%, expressed that the most important content taught to them had been interpreting and translating career-specific texts from English into Spanish, curiously, the one which was not in the curriculum.

References

Daniels, H. (Ed.). (2005). *An introduction to Vygotsky*. New York: Psychology Press.

Dudley-Evans, T., & St. John, M. (1998). *Developments in English for specific purposes: A multi-disciplinary approach.* Cambridge: Cambridge University Press.

Gigena, A. (2010). El componente motivacional en la diagramación de desarrollos curriculares de lengua extranjera para nivel universitario. In *Formación e investigación en lenguas extranjeras y traducción. Actas de las segundas jornadas internacionales*. Pp 429-433. Retrieved from: https://www.academia.edu/13777626/El_componente_motivacional_en_la_diagramac iC3%B3n_de_desarrollos_curriculares_de_lengua_extranjera_para_nivel_universitari o

Pulverness, A. (2017). Squaring the circle: Teaching EAP to large mixed groups. *The English Teacher*. Retrieved from: http://journals.melta.org.my/index.php/tet/article/view/353

Knight, P (2001). Complexity and curriculum: A process approach to curriculum-making. *Teaching in Higher Education, 6*(3), 368-381.

Robinson, P. (1991). *ESP today: A practitioner's guide.* London: Prentice Hall.

Book review

**Children and teachers as co-researchers:
A handbook of activities**

Annamaria Pinter and Rama Mathew
London: British Council 2016, 102pp
ISBN: 978-0-86355-856-6, PDF, free download at
https://englishagenda.britishcouncil.org/sites/
default/files/attachments/27780_eltra_paper_final.pdf

Teaching English as a Foreign Language (TEFL), as well as teaching any other subject, is a challenge considering the changing contexts in which educators are immersed nowadays. In this book, the editors, professors and developers in the educational area, Annamaria Pinter and Rama Mathew, seek to offer a set of real-life activities that were implemented by seventeen Indian teachers on a previous project called "Children and teachers as co-researchers in Indian primary English Classrooms," which began officially in February 2015.

The editors of this handbook were Annamaria Pinter and Rama Mathew. The former is an Associate Professor at the centre of Applied Linguistics in the University of Warwick, who has published widely in the area of teaching English to young learners. The latter is a retired Professor of Education from the University of Delhi, who has been involved in several teacher development initiatives, including 'Portfolio Assessment' and 'Diary Writing' within a teacher-as-researcher framework.

As it is described in the introductory section, the aim of this handbook is to share a wide set of activities derived from the project mentioned before, in which children have expressed what they would like to do in class and have collaborated with their teachers exploring aspects of their own learning. This volume includes an introduction and six sections that contain forty-five activities split into different themes.

With a structure which is easy to follow, filled with pictures and charts, this handbook attracts a wide range of readers. It is useful not only to English language teachers, but also to
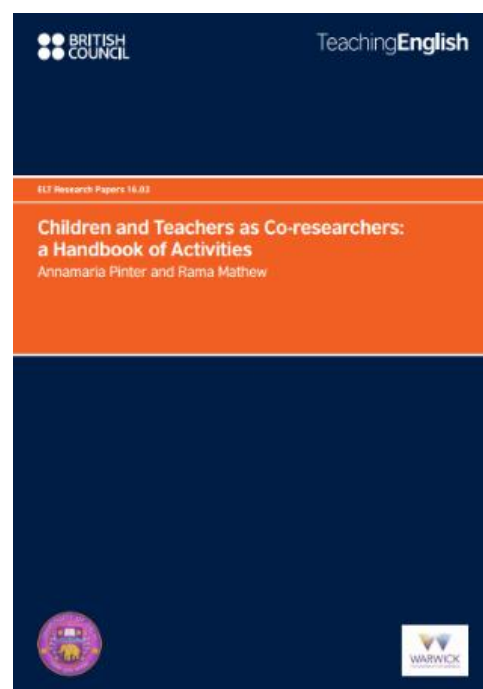
those who are interested in research and in changing the focus of lessons for students to be the centre of their own learning process. The language used throughout the book is clear and suitable for its purpose. The information included is concise and straightforward and the necessary explanations about the project are provided at the beginning.

As a way of guiding the readers, the introduction offers a sub-section based on the project from which the activities emerged, including the context and factors to be considered before reading, explaining the different types of schools with which Indian teachers worked and, furthermore, how the activities worked as outlines for children to be co-researchers. Moreover, the introduction has another section about the handbook, which gives an overview of its structure and content.

Each of the six following sections are numbered and describe a particular topic, such as 'Trying out something new,' 'Children making choices,' 'Handing over some control to children,' 'Building positive relationships,' 'Designing research tools and analysing empirical data,' and 'Feedback and self/peer- assessment.' All the sections have several activities with the name of the teacher who carried them out, the age of the students, the materials needed, the organisation and, in some cases, a description. Furthermore, most of the activities are presented together with charts, and at the end there is a comment from the editors which is useful to understand the purpose of the activity and how it helped both teachers and students. While the examples of the activities provided are of great help to visualize and afterwards put the activities into practice, there is no specific information about the teachers nor the different groups of students, only their age.

Overall, this volume contains a wide range of ideas to adapt and put into practice in different classroom situations, depending on the context of the teachers and the students. Even though there is no conclusion from the contributors, one can perceive a sense of positivity towards this way of working *with* students more than *for* students.

As an educator, I consider that the contributors succeeded in developing different tools within their classrooms in order to improve children's autonomy and their relationship among themselves and with the teachers. Also, the editors achieved the aim of communicating effectively the benefits of having children and contributors working together with examples taken from real-life experiences. I would highly recommend this handbook to all the educators who are interested in research as a tool for improving their classroom outcomes and for taking into account their students' voices and feelings. This book clearly states the advantage of children getting involved in their own learning, which is that they are much more motivated and learn more, which is important.

Teachers in other contexts would surely benefit from reading this book as well as from including adaptations of the activities in their own classrooms in order to experience a change in their way of teaching and the process of learning. Moreover, students worldwide can benefit from engaging in a meaningful dialogue that helps them to become co-researchers and to identify their learning process more significantly.

Agustina Paniagua
ISFD N° 809, Esquel
agus.paniagua.19@gmail.com

Book review

A handbook for exploratory action research

Richard Smith and Paula Rebolledo, London, British
Council, 2018, Pp. 115, ISBN 9780863558856

The aim of *A handbook for exploratory action research*
is to introduce teachers to the personal and professional
benefits of researching their own classrooms for
continuing professional development and for students'
improvement in learning a second language. In other
words, the main purpose is to help teachers and their
students achieve greater success in the classroom.

The book is written in a jargon-free, non-
academic style to achieve a reader friendly form of
presentation that will be appealing rather than off-
putting. This characteristic makes primary and secondary school teachers or novice teachers
in the action research field the target readership. The book proposes an approach that enables
teachers to gain a better understanding of their classroom contexts and thus develop more
appropriate ways of teaching. The title of the book is based on a programme that has already
shown its worth in India and Nepal, which are also developing countries as Chile and
Argentina. The aim of the programme is the promotion of intercultural understanding of
teacher-researchers around the world and the building of a sense of a global ELT teacher-
research community. The fact that the authors have different nationalities is also a clear
intercultural example in this work.

The book is organised into an introduction, nine chapters and a section with extra
materials. In the introduction the authors present the topic of the book creating a space for
reflection and exploration in order to understand, analyse and solve difficult situations
through exploratory action research. The idea is that teachers facing problems such as heavy
teaching loads, large classes or lack of material resources can understand these situations
better before taking action.
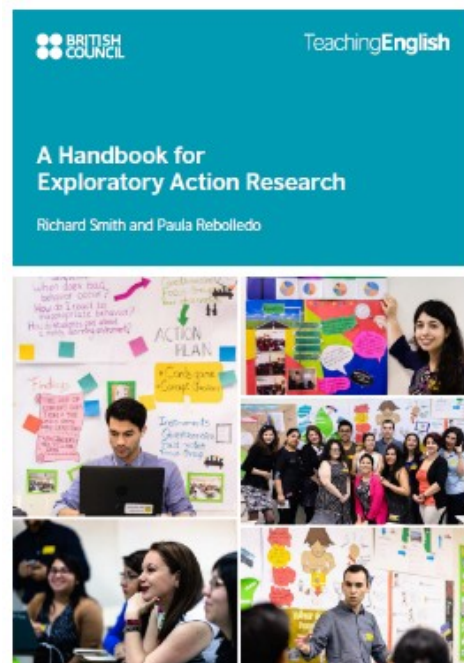
Chapter 1 considers the general characteristics of research and invites readers to reflect on a situation in the classroom and to identify signs of success, or lack of it, to understand it as a first step to doing research. The value of collecting information is that teachers gain new perspectives that can be useful to solve a problem or a puzzling situation.

Chapter 2 tries to demystify research and stresses the idea that research is not something only scientists do. Teachers themselves can carry out research in issues of importance for them in their own classroom. The whole process of research involves seven steps explained in detail in the following chapters and with real examples.

Chapter 3 looks in more detail what is involved when a teacher explores his or her own practice. In this section the authors develop the stages in the process of explanatory research which imply: clarifying your problem, interest or puzzle, and deciding on your specific questions; collecting the data that you need to answer the questions and, finally, analyzing and interpreting the data to answer the questions and better understand your situation.

Chapter 4 deals with the motivations for doing research and how to choose a particular topic. The extent to which each possible topic for research should be "Manageable", "Urgent", "Significant" and "Engaging". Then, it explains step by step how to make research questions as smart as they can be. Next, to answer these exploratory questions Chapter 5 suggests the process of "triangulation", that is the use of different methods to look for external sources of information. After obtaining the answers, Chapter 6 explains how to analyse and interpret qualitative and quantitative data collected from the teachers' findings.

In Chapter 7, on the basis of the new understanding gained through exploration, we come to the action phase: how exploratory research can turn into action research if some kind of new action is needed. Chapter 8 focuses on observing and analyzing the effects of that new action introduced in the action research part of the project. Finally, Chapter 9 invites teachers to join wider communities such as IATEFL and TESOL and become active members of a global ELT teacher-research community. There is also a final chapter dedicated to extra materials. Teachers can find a sample questionnaire and an observation checklist together with an Answer Key for the many practical tasks.

The approach presented in the book is based on a programme that is currently in its fifth year in Chile. While particularly situated in Chile and appropriate to the Chilean experience, using concrete examples, the book appeals for novice teachers or teachers facing difficulties like large size classes, lack of preparation time, lack of resources, and students' demotivation to learn English.

Reading real stories of teachers facing problems that we may have also coped with, or will have to cope with in the future is really valuable and helpful to novice teachers. It is a way of having a guide on how to find solutions since the handbook provides practical activities and reflective tasks with clear instructions. Consequently, *A handbook for exploratory action research* is designed to be used by *any* teacher, on their own, with a

colleague or colleagues, or within teacher associations. As the book is written by teachers who know that visuals support understanding, the use of colourful illustrations and photographs enhances its appeal and accessibility even further.

Personally, what really helps me empathise with the situations and teachers' experiences presented is the reflection on my own experience through the practical tasks proposed. The extra reading activities provided in each chapter present other situations experienced by teachers in order to illustrate theoretical aspects that make us reflect on our practice.

Research is of fundamental importance to our professional development. This can be clearly seen in the teachers' findings and how their actions have truly improved their teaching strategies. Doing research is more like an invitation to a journey of exploration and action, rather than an academic task. This would also be the beginning of a longer process of development and interaction with other teacher-researchers in a context of English as an international language. Implementing new ideas in the classroom enables us to develop our own creativity and our critical analysis to make adaptations if necessary.

To conclude, I highly recommend this book, because, as teachers, it is of paramount importance that we reflect on our own practices in order to improve them. The approach presented in the book was so successful that the Ministry of Education in Chile is planning to distribute this book to secondary teachers in Chile. I think this action would be fruitful: the more teachers can explore their own classrooms, the more changes for the benefits of the students would be implemented.

Cintia Vanesa Alderete
Profesorado Superior Lenguas Vivas Nº6007, Salta
alderete.vanesa@yahoo.com.ar

Book review

Developing as an EFL researcher: Stories from the field

Developing as an EFL researcher: Stories from the field can be considered the conference proceedings of the 2015 IATEFL ReSIG Pre-Conference Event, which took place in Manchester, England. The conference "Developing as a Researcher" consisted of five-minute presentations of posters created by the speakers, and the e-book is a selection of papers developed by some of the participants from the posters presented at that meeting. Both the conference and the book were conceived for sharing stories of personal growth and perseverance around becoming a researcher in TESOL/EFL area, and spreading novel perspectives around the nature of qualitative research. Significantly, the threads that run throughout the papers are the role of subjectivity in qualitative research and the contribution of the collective action of the ReSIG to the professional development of researchers.

The resulting collection is a myriad of narratives of an emotional tone about the efforts, sacrifices and satisfactions associated with research practice. The affective content is fairly balanced with thoughtful instances of reflection on methodological aspects such as the steps in the process, a few ethical concerns and several refreshing hypotheses about subjectivity, identity and context in research that may lead to further exploration beyond the action sphere of the ReSIG. This book offers not only an emotional and careful self-reflective approach to the field of research in EFL, but also a valuable contribution to keeping the flame of vocation burning in the heart of practitioner-researchers and awaking in learners and teachers the interest for research.

The editorial work, in charge of Sian Etherington and Mark Daubrey, organized the book around three strands of interest and some key ideas covered during the panel discussion at the end of the conference in Manchester. The whole selection of papers is preceded by an acknowledgement section, a note on the contributors with information about their credentials, and an introduction in charge of Sian Etherington that includes a succinct reference to the conference events, an outline of the texts included in the book and an editor's personal reflection upon the encounter in Manchester. This digital edition comes with features of interactive nature. For instance, each paper includes a digital link to a YouTube video of the poster presentation, and an active "Contents" icon facilitates the navigation to specific sections of the book.

Each strand is introduced by the paper of the impulse speaker, an invited discussant of significant trajectory in the field of EFL such as David Nunan, Sue Garton and Cynthia White. The first strand, "A researcher's journey: Challenges, issues and strategies," is introduced by Nunan's paper in which the author shares a shocking experience that marked the beginning for his research activity. It is followed by four participants' narratives of their own paths along the route of research. The second strand, "Specific methods and specific challenges," consists of four papers into key features of qualitative research, the inherent messiness to qualitative data and the need for openness and tolerance to complexity as an indispensable methodological disposition. The third strand, "Identities, roles, relationships and contexts in research," represents the bulk of the book and includes six papers dealing with the central role of researchers' and participants' subjectivities in every stage of the research process.

Each strand of the book offers a new perspective on the need of examining more seriously the role of affective factors in qualitative research and the significance of transforming the practice endeavour into a shared experience of learning. The book considers an often-neglected truth: at the end of the day, personal and professional life are deeply intertwined and both are motivated by the promise of improvement.

Strand 1 includes multiple pearls of wisdom that beyond the specific circumstances that helped them to flourish can help many others to face the challenges the research journey may carry in a multiplicity of contexts. Becky Steven's emphasis on the importance of sharing findings with partners, Irena Mestrovic's invitation to discuss more openly the affective dynamics of being a researcher, Assia Slimani-Rolls's findings about the impact of teachers' engagement with research on their own professional development, or Alexa Piaggio's exercise of intellectual honesty in the face of failure will offer several instances of deep reflection.

Strand 2 focuses on the specific challenge of doing qualitative research and classroom-based research. Issues such as the dilemma within the researcher-participant relationship, the definition of effective data collection, the tension between real-life messiness and theoretically informed interpretations, the marginalization of the participants'

voice, the need of growing as a disciplined writer, as well as the benefits of sharing the journey, will provide the reader with a vivid picture of the forces defining the dynamics in different stages of the practice and the possibilities for further systematization.

Strand 3 concentrates on the person of the researcher affected by the complexity of the context of work. Considerations around personal trajectories, such as playing contradictory roles in the classroom, dealing with dilemmas or the complexities within the implementation of a teacher education programme, will bring to light the centrality of the researcher's subjectivity and how their identity is shaped by a context of shifting relationships and the need to fulfil multiple roles.

The book ends with a summary of the ideas that emerged during the panel discussion, few final remarks by Mark Daubrey, and a final invitation to become a member of the ReSIG.

Developing as an EFL Researcher captures real life experiences and honest reflections on the nature of the research practice by presenters from different cultural backgrounds, all shared in the context of the supportive and fruitful environment provided by the ReSIG. It will both inspire professionals to persevere in their work and awake in outsiders the desire to explore the field and consider research as a fundamental aspect of their development as teachers. This book may inspire other groups of professionals to build community sense through the sharing of experiences around the topic and to inquiry further in the role of classroom-based research as a tool for professional development.

Jimena Ferrario
Profesorado Superior de Lenguas Vivas N° 6007, Salta
jimena_ferrario@hotmail.com

Book review

International perspectives on teaching the four skills in ELT: Listening, speaking, reading, writing

Edited by A. Burns and J. Siegel, Basingstoke, Palgrave Macmillan, 2018, Pp. 255, ISBN 978-3-319-63444-9 (e-book)

International Perspectives on Teaching the Four Skills in ELT: Listening, Speaking, Reading, Writing compiled by Anne Burns and Joseph Siegel is a fascinating collection of testimonies from English teachers all over the world and their experiences on teaching the four skills. As it provides such a broad scope of experiences, this volume can be useful and compelling not just for novice teachers but also experienced ones.

The book opens with an introductory chapter written by the editors Anne Burns and Joseph Siegel in which they explain the themes and issues that come up when teaching the four skills. Furthermore, the volume is divided into four parts, one for each one of the skills. Each chapter is in turn divided into sub sections such as Introduction, Implications, Conclusion, Questions for Reflection and a Reference section.

Part 1 delves deep into listening and it is four chapters long, going from Chapter 2 to 5. Chapter 2 is written by Denise Santos and Suzanne Graham and features a striking contrast in the listening pedagogy in an as a Foreign Language (EFL) classroom in Brazil and an English as a Second Language (ESL) classroom in the English speaking UK. The third chapter, written by Willy A. Renandya and Guangwei Hu deals with an examination of the way L2 listening is currently being taught in China. The fourth chapter, which was written by Michael McAuliffe and Gavin Brooks discusses the development of a listening course inside the mostly monolingual Japanese universities. The fifth and final chapter is written by M. Gregory Tweedie and Robert C. Johnson and explores the uses listening can
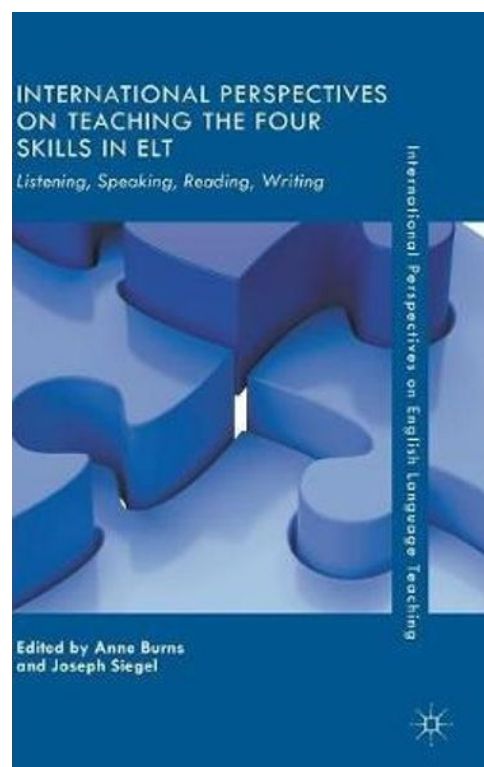
have in English for Specific Purposes (ESP) more specifically in the context of nursing education which is where they conducted their research and where English is commonly used as a Lingua Franca.

Part Two encompasses Chapters 6-9 and deals with the pedagogy of speaking. Chapter 6 deals with teaching speaking skills in the context of primary schools in Cameroon. It is written by Achu Charles Tante and mentions the ever so frequent problem ESL and EFL teachers have: just how complicated it is for students to express themselves in English. Chapter 7, written by Phillip Chappel, explores inquiry dialogue as a genre to promote teacher and students speaking skills in the classroom; chapter eight, written by Winnie Pang and Michael Burri, dives into the uses dialogic speaking skills can have in English for Academic Purposes (EAP) classes in Canada. Olga Kozar wrote the ninth and final chapter of the speaking section and it features a Russian perspective on using Skype as a tool to teach conversational English to adults. This chapter can be intriguing to teachers all across the globe since technology has come to stay in the field of education.

Part 3 of this volume is dedicated to reading and it encompasses Chapters 10-13. Chapter 10 is written by Olga Vrastilova. It explores the possibility of using authentic literary texts in elementary school classes. As she mentions in this chapter, reading is often adapted in texts aimed at EFL students but she argues that elementary school students can actually access and understand authentic material, such as children's literature written in English. Sri Rejeki Murtiningsih and Winda Hapsari write in Chapter 11 about how reading can be used to further develop critical thinking skills and collaborative work. In Chapter 12, Gordon Blaine West argues that autonomy can also be developed in the context of an ESL academic reading course through the reading skill. Finally, in Chapter 13, the last chapter of the reading section, Kevin Roach deals with the social significance reading can have in adult migrants. His captivating argument is that, in order to teach reading skills to adult migrants, teachers need to be aware of literacy as a social practice.

Finally, Part 4 is about writing and it encompasses Chapters 14-17. Chapter 14 is written by Rawia Hayik and is about promoting descriptive writing through culturally relevant literature in the context of Israeli college students whose first language is Arabic. Nhu Luan Pham and Noriko Iwashita delve into using corrective feedback in order to enhance Vietnamese learner's autonomy in Chapter 15. In Chapter 16, Ricky Lam explores promoting self-reflection in writing through the use of showcase portfolios. This section concludes in Chapter 17 which is written by Brazilian teacher Isabela Villa Boas who argues for localizing the pedagogy of second language writing through a skills-integrated language program in order to create a more meaningful written production. Once the writing section is over, the volume has a final concluding chapter which, like the introductory chapter, was also written by the editors Anne Burns and Joseph Siegel. In this final chapter –Chapter 18- the authors theorize about the possibilities that the four skills might take in the future.

Overall, because of its broad scope, *International Perspectives on Teaching the Four Skills in ELT: Listening, Speaking, Reading, Writing* is a compelling read for both novice English teachers and experienced ones. That being said, it does not mean that there are not chapters that feel somewhat lagging or those that present issues that the reader might not find interesting or relatable. Nevertheless, this volume is highly recommendable and a good read. Personally, I agree with Burns and Siegel captivating argument in their concluding Chapter 18 where they state that, as English speakers interact in a more and more globalized community outside the classroom, they will need to integrate the four skills in much more complex ways than the ones they are currently sitting through at school. Consequently, it is time for teachers to rethink the way we are teaching the four skills to our students because seeing the four skills as completely separated entities is no longer viable and the English curricula all around the world needs to be modified in order to suit the needs of these highly globalized students.

Carla Florencia Lescano
Universidad Católica de Salta, Salta
c.lescanof@gmail.com

Book review

A handbook for exploratory action research

R. Smith and P. Rebolledo, London, UK: British
Council, 2018, Pp. 115, ISBN 978 0 86355 885 6
This publication is free to download in pdf format.

A Handbook for Exploratory Action Research is a
detailed manual written in a way which is inspirational
for a wide readership of primary, secondary school
teachers and teacher educators. Comprehensive in its
coverage and moving from real examples taken from the
Chilean Champion Teachers programme, this handbook
focuses on the value and the benefits of teacher research
on their own practice in their teaching contexts.
Throughout the book, Smith and Rebolledo emphasize
the importance of Exploratory Action Research
(hereafter EAR), that is to say, teacher exploration in their own classrooms by means of data
collection, analysis and subsequent reflection on the possibility of introducing change.

The authors begin the handbook with an Introduction before presenting nine chapters,
which are all introduced by questions related to this kind of practitioner research, and
Chapter 10 which provides not only extra material but also a detailed answer key section. In
the Introduction after EAR is defined and the way in which it was developed is explained,
characteristics of the style of presentation of the handbook and ways of using it are given.
The introductory section concludes with acknowledgements and a self-diagnosis activity
which is repeated at the end of the book.

Each of the following nine chapters is divided into a series of headings that relate to
real teacher stories from the Champion Teachers programme, tasks to be solved by the
readers most of which can be checked in the Answer Key section in the back of the
handbook, a summary and a follow-up section. Chapter 1 introduces readers to the concepts
of 'success stories' and 'challenging teaching situations' and makes them reflect on their
own teacher practice. 'Sign' and 'evidence' with relation to research are also explored in this

chapter and the importance of research is stressed through the story told by a Chilean teacher. Chapter 2 deals with the significance of teacher research, a pair of misunderstandings associated with it and its differences with traditional research.

Chapter 3 addresses the value of exploration and the stages which constitute exploratory research and action research. The way to select the topic of research and how to formulate the exploratory questions are dealt with in Chapter 4. Chapter 5 introduces several sources of information and emphasizes how significant it is that the methods adopted by the teacher researcher match effectively the research questions. The conception of data analysis is examined in Chapter 6 and ways of analysing and making sense of qualitative and quantitative data are covered.

In Chapter 7 the possibility of engaging in action research is presented and readers are encouraged to go through the different stages of that process. Through several Champion Teachers stories in Chapter 8, different data collection methods and ways of interpreting that data critically are reviewed. The importance of triangulation is emphasized once more and the central changes in 'mindset' are also mentioned. In Chapter 9, the authors consider the question: Where do I go from here? They begin their response by making readers reflect on whether this is the beginning or the end of their journey and move on to describe different ways in which they can share their EAR findings. Finally, the extra material presented in Chapter 10 is brief but useful. For those novice teacher researchers the Answer Key section at the end of the chapter will be an enormous resource.

The clarity with which this handbook has been written and organised makes it a very user-friendly resource for all teachers, especially for those who are beginning to implement EAR even in low-resourced contexts. By emphasizing the importance of teacher research, this book makes a positive and practical contribution to a field which was unexplored by many readers as it focuses on teachers working in their own classrooms, generating practical knowledge after exploring their own unasked questions, making changes if necessary and sharing the outcomes of their research plans. In *A Handbook for Exploratory Action Research,* novice teacher researchers will find new insights, teachers' accounts from their own experiences, commentaries, tasks and new motivations.

Mariana Serra
Instituto Superior de Formación Docente y Técnica N°125,
Buenos Aires
marianaserra2@gmail.com

Guidelines to authors

Please, read carefully all the sections below before submitting your manuscript. The Editor may reject your submission if your manuscript does not follow the instructions below.

To publish with AJAL, all Argentinian authors or half of them in a multi-authored article must be members of a teacher association (TA) linked to FAAPI by the time the submission is made. If you are the sole author of the manuscript submitted, please ensure that you have become a member of a (TA) before submitting your manuscript. **Make sure that your contribution, even when based on a local context and issue, reaches an international readership.**

International authors who are the sole authors of the manuscript submitted do not need to be part of a TA in Argentina for publishing with AJAL. If you are an international author and your manuscript is co-authored by an Argentinian author, please ensure that the latter is a member of a TA linked to FAAPI. If you are an international author and your manuscript includes other international co-authors, please contact the Editor at ajaleditor@faapi.org.ar before making your submission. The Editors will ensure that each issue has a balanced representation of Argentinian and international authors.

AJAL does not accept articles authored by more than four authors.

Ethics and copyright

By submitting your manuscript to AJAL, you confirm that you have read and followed all instructions to authors. You confirm that it is an original manuscript whose authorship belongs to you and that it has not been published previously or under consideration for publication elsewhere. By submitting your manuscript you also confirm that all potential conflicts of interest have been taken care of successfully and that your manuscript does follow strict ethical procedures such as data protection, anonymity, consent forms from participants, permissions from published materials or any third parties. Please, see the author form which you then must complete and submit together with your manuscript.

Once your article is accepted, you should authorise AJAL to reproduce the full text on AJAL website. This will imply filling in a publication license form to be indicated by the editor. If you wish to use the article again in a publication written or edited by you, you may do so provided that its original publication in AJAL is acknowledged.

Types of manuscripts you may submit
We welcome the following types of manuscripts:

Original research article: this is a manuscript which reports the outcomes of your own research. It usually includes sections such as introduction, literature review, methods, results,

discussion, conclusions and implications, and references. Length: between 5,000 and 7,000 words (including tables and graphs, without references and appendices).

Literature review/state-of-the-art article: this is a critical review of the literature on a particular theme of interest to AJAL. Sections may vary but authors are encouraged to include conclusions which promote reflection and suggestions for further research. The article should contain a minimum of 30 quality references. Length: between 3,000 and 5,000 words (including tables and graphs, without references)

Reflective article: this is an article based on theoretical underpinnings. It includes description, analysis, and evaluation of a specific topic or issue. Length: between 2,000 and 3,500 words (including tables and graphs, without references).

Interview-based article: this is an article which consists of an interview with a specialist on a specific field within applied linguistics and it includes a framework of reference and concluding remarks. Length: between 3,000 and 4,500 words (including tables and graphs, without references).

Classroom activities: this is an article which describes strategies, techniques, or activities which could be applied to the EFL classroom. The Editors will favour those which are the product of your own creation and/or adaptation and that you may have tried. The manuscript should be highly practical with clear instructions, and direct classroom application or implementation. Length: between 1,000 and 3,500 words (including tables, without references and appendices).

Classroom account: this is a narrative article based on personal classroom experiences such as projects, materials development, or other special systematic undertakings. Length: between 1,000 and 3,500 words (including tables, without references and appendices).

Materials review: this is a review of a book, software, website, or any other resource of relevance to AJAL readers. Length: between 600 and 1,500 words. Before submitting your manuscript, please contact the Materials Review Editors at ajalmatreviews@faapi.org.ar to discuss the material you wish to review.

How to prepare your submission
- Read these instructions carefully. For issues which are not included here, please contact the Editor at ajaleditor@faapi.org.ar before submitting your manuscript.
- Authors must submit a complete manuscript. We do not accept abstracts prior to a complete submission.

- Do not format your submission in the style of a published article. All manuscripts should be sent as Word documents containing unjustified texts.
- Authors' names, self-referentials (e.g. As I have outlined elsewhere (Wright, 2011)) and affiliations MUST NOT appear in the manuscript. Remember that all manuscripts undergo blind reviewing.
- Make sure that your manuscript is written in accurate English (AmE and BrE spelling are accepted but these should not be used simultaneously). Please, have your manuscript proofread or edited before submitting. Manuscripts in need of serious editing will be rejected and will not go through the peer review process.
- Make sure you respect the word count according to the type of manuscript you submit.
- Your manuscript should be sent as a Word document with the following format: Times New Roman, 12, double-spaced, unjustified). Do not insert page breaks in your manuscript.
- With the exception of materials reviews, all manuscripts should include a title. The title must be clear and self-contained. Please avoid long titles. Only capitalise the first word and proper nouns.
- The title of your materials review should start like this: 'Review of (title of book or website) by (authors). If it is a book, also include the information (if applicable) as shown in the example below:

  > **CLIL. Content and Language Integrated Learning**
  > D. Coyle, P. Hood and D. Marsh
  > Cambridge
  > Cambridge University Press
  > 2010
  > Pp. v + 173
  > ISBN 978-0-521-11298-7 (hbk): £54.50; US$ 71.20
  > ISBN 978-0-521-13021-9 (pbk): £21.00; US$ 25.91
  > ISBN 13:9780511740473 (Adobe e-reader): US $71.00

- With the exception of materials reviews, all manuscripts should have two abstracts, one in English and one in Spanish. Each abstract should be around 100-120 words. The abstracts must be followed by a maximum of five key words in both English and Spanish.
- Indent all paragraphs except the first paragraph of each section.
- No line space should be left between paragraphs or under (sub)headings. Line space should be left between sections.
- Avoid the use of footnotes. If they are necessary, place them after your conclusions as endnotes.

- Use double inverted commas for short quotations and single inverted commas for quotations within quotations.
- To highlight a word or a concept, use *italics*.
- Indent long quotations (40 words or longer).
- For all quotations refer to authors as follows:

> According to Levin (2010, p. 359), 'governments around the world continue to be intensively involved in changing their education systems.' (for long quotes you may place the author's surname, year: page sequence below the quote, ranged right)

- For lists use Arabic numerals.
- For bullet-points, use □
- Figures and tables should be clearly labelled with a number and caption. For example: Figure 1. Types of motivated behaviours. (For captions use Times New Roman, 9). Format your figures and tables as you wish them to appear.
- For materials reviews, do not include appendices or use headings.
- With the exception of materials reviews, appendices must be signalled in the text and then placed after your reference list. Label appendices as Appendix A, B, C…
- With the exception of materials reviews, use headings and subheadings. Please, do not name the first section of your manuscript. Name the sections in which your manuscript is divided following the example below:

> **Methodology** (Times New Roman, 12, bold face, indented, upper case and lower case headings).
> **Data Collection Instruments** (Times New Roman, 12, left-aligned, uppercase and lowercase headings)
> **Interviews** (Times New Roman, 12, bold face, indented, a period, lowercase heading).

- For in-text references follow these examples:

James (2009) argues that…

> Gómez and Pérez (2008) raise other issues since…
> The situation in Argentina has shown relatively low improvement (Andes, 1998; Gómez & Pérez, 2008; Zander, 2000).
> Little (2006a) observes that…
> Little (2006b) denies that…

- For works authored by three or more authors, include all surnames the first time you refer to them, and et al. in subsequent references, for example:

> Smith et al. (2010) signal that…
> This has been signalled by many works (Smith et al., 2010)

- Full references: all authors cited in your manuscript must appear in your reference list. Follow these examples:

Bruner, J. (2002). *Making stories: Law, literature and life.* Cambridge, MA: Harvard University Press.

Gee, J.P. (2005). Semiotic social spaces and affinity spaces: from the Age of Mythology to today's schools. In D. Barton & K. Tusting (Eds.), *Beyond communities of practice* (pp.214-232). Cambridge: Cambridge University Press.

Lantolf, J. (Ed.). (2000). *Sociocultural theory and second language learning*. Oxford: Oxford University Press.

Little, D. (1991). *Learner autonomy 1: Definitions, issues and problems*. Dublin: Authentik.

Meza Rueda, J.L. (2008). *Historia de maestros para maestros. Pedagogía narrativa expresada en relatos de vida*. Bogotá: Universidad de La Salle.

Pérez-Cañado, M.L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism, 15*(3), 315-341.

Prince, P. (2011). What's the story? Motivating e-learners with fiction. In D. Gardner (Ed.), *Fostering autonomy in language learning* (pp. 225-233). Gaziantep: Zirve University.

Richards, K. (2006). "Being the teacher": Identity and classroom conversation. *Applied Linguistics*, *27*(1), 51-77.

Smith, R. (2003). Teacher education for teacher-learner autonomy. In J. Gollin, G. Ferguson & H. Trappes-Lomax (Eds.)*, Symposium for language teacher educators: Papers from Three IALS Symposia.* Edinburgh: IALS, University of Edinburgh. Retrieved May 2, 2012, from http://homepages.warwick.ac.uk/~elsdr/Teacher_autonomy.pdf

Smith, R., & Erdogan, S. (2008). Teacher-learner autonomy: Programme goals and student-teacher constructs. In T. Lamb & H. Reindeers (Eds.), *Learner and teacher autonomy: Concepts, realities and responses* (pp. 83-102). Philadelphia, PA: John Benjamins Publishing Company.

Ushioda, E. (2011). Motivating learners to speak as themselves. In G. Murray, G. Xuesong & T. Lamb (Eds.), *Identity, motivation and autonomy in language learning* (pp.11-24). Bristol: Multilingual Matters.

What to submit and how

1. You must submit the following documents:
   a. Author form
   b. Complete manuscript in Word format (including tables and figures).
   c. If applicable, you must submit tables and figures as separate files: submit tables as Word documents and figures/illustrations in TIFF format.
2. **With the exception of materials reviews, all other submissions (documents a-c above) must be sent to ajaleditor@faapi.org.ar**
3. Materials reviews (documents a-b above apply here too) should be sent to **ajalmatreviews@faapi.org.ar**

What happens once you submit your manuscript

1. You will receive an email acknowledging receipt in around 4 days.
2. We expect to return to you with the evaluators' comments in around 40-60 days.
3. If your manuscript is accepted with minor changes, you will be expected to resubmit your manuscript in 15 days.
4. If your manuscript is considered for 'revise and resubmit' (major corrections), you will be expected to resubmit your manuscript according to a time frame agreed with the editor.

AJAL