AELINCO 2015
BOOK OF ABSTRACTS

7th Conference on Corpus Linguistics Valladolid (Spain)
5-7 March 2015
Organizing Committee

Pedro A. Fuertes Olivera (Chair)
Esther Álvarez de la Fuente
Raquel Fernández Fuertes
Pilar Garcés García
Belén López Arroyo
Marta Niño Amo
Isabel Pizarro Sánchez
Ana Sáez Hidalgo
Ángeles Sastre Ruano
Marisol Velasco Sacristán

Student Helpers

Tamara Gómez Carrero
Noelia Recio
Silvia Sánchez Calderón

Scientific Committee

Francisco Alonso Almeida (Spain)
Theo Bothma (South Africa)
Pascual Cantos Gómez (Spain)
Gloria Corpas Pastor (Spain)
Raquel Criado Sánchez (Spain)
Danie Prinsloo (South Africa)
Teresa Fanego Lema (Spain)
María Luz Gil Salom (Spain)
María de los Ángeles Gómez González (Spain)
Sylviane Granger (Belgium)
Andrew Hardie (UK)
Ulrich Heid (Germany)
Julia Lavid López (Spain)
María José López Couso (Spain)
Juana I. Marín Arrese (Spain)
Antonio Moreno Sandoval (Spain)
Isabel Moskowich-Spiegel Fandiño (Spain)
José Luis Oncins (Spain)
Javier Pérez Guerra (Spain)
Emilio Ridruejo Alonso (Spain)
Jesús Romero Trillo (Spain)
Chelo Vargas Sierra (Spain)

And the members of the AELINCO Board

Aquilino Sánchez Pérez (Spain)
Mª Luisa Carrió Pastor (Spain)
Miguel Fuster Márquez (Spain)
Antonio Moreno Ortiz (Spain)

Sponsors

Financial support for this event has been received from

Grant no FFI2011-22885: Ministerio de Economía y Competitividad; Principal Investigator: Pedro A. Fuertes Olivera

Grant no VA067A12-1: Dirección General de Universidades e Investigación; Junta de Castilla y León; Principal Investigator: Pedro A. Fuertes Olivera

Institutional Sponsors

Universidad de Valladolid:
Departamento de Filología Inglesa
Research Unit: International Centre for Lexicography
Ministerio de Economía y Competitividad
Dear AELINCO participants,

Last year in Las Palmas we invited you to Valladolid. Now we would like to welcome you warmly to the 7th International Conference on Corpus Linguistics (5-7, March 2015) and to our old and beautiful city in hopes that you will profit from the very promising conference sessions. This conference will confirm that corpus linguistics is at a crossroads and will enable possible future ways of advancing, either by accepting the coming of big data or by denying its relevance. Arguments on both sides will be discussed providing us with inspiring and stimulating debates on the subject.

We also encourage you to find some spare time here and there and do some sightseeing: to enjoy the city cozy places, our food (e.g. tapas) and wine, and the friendly and lively atmosphere of Valladolid. We do hope that the weather will be “merciful” to us, but we should not forget that we are approaching spring time, when weather changes are as common as the old Spanish adagio: “marzo marzuelo, un día malo y otro bueno” which reminds us that in March the weather shifts, from bad to good in days or even in hours.

The AELINCO organizing committee

Pedro A. Fuertes Olivera, Esther Álvarez de la Fuente, Raquel Fernández Fuertes, Pilar Garcés García, Belén López Arroyo, Marta Niño Amo, Isabel Pizarro Sánchez, Ana Sáez Hidalgo, Ángeles Sastre Ruano, Marisol Velasco Sacristán
Plenary Speakers

Andrew Hardie (University of Lancaster, The UK): Big data in language studies: from cargo-cult science to phantom revolution

“Big data” has recently become a buzzword in a range of academic disciplines as well as in certain quarters of business and government, although the criteria by which data counts as “big” are not always clear. The standard account is that the age of big data, which arrived approximately three to four years ago, promises revolutionary change across all the fields where it can be applied. What does this mean for the study of language and text? In this lecture, I will argue that for linguistics, work that can be labelled as using “big data” falls under two banners: that which is novel is not good, and that which is good is not novel. First, I will look at some of the fruits to date of the big data approach to language usage data. The last several years have seen numerous high-profile academic publications – albeit not in linguistics journals – purporting to implement such an approach. I will explore some examples of such studies, illustrating basic flaws which prevent them from rising above the level of pseudo-research. Second, I will develop a critical appraisal of the concept of “big data” in the textual sense: what does the term even mean in the context of language and linguistics? Thus, I will argue that the prospect of a big data revolution in linguistics is fundamentally illusory.

Paul Rayson (University of Lancaster, The UK): Big Data Challenges with Big Corpora and Big Taxonomies

Over the last forty years, corpora have grown from small collections and representative datasets of around one million words, through national corpora of around 100 million words, to vast collections of billions of words. With new web-derived or online social media datasets plus mega corpora resulting from national and international digitisation initiatives, the trend for corpora is ever increasing size and breadth. Over this same time period, successive generations of corpus annotation and retrieval software have increased in power and speed, although much the same retrieval methods (frequency lists, concordances, keywords, collocations and n-grams) are still applied. In this talk, I will present our work in the Samuels project (www.gla.ac.uk/samuels/) where we are studying two very large corpora of the order of 1-2 billion words each: (a) Early English Books Online (EEBO) Text Creation Partnership (TCP) consisting of over 48,000 books published between 1473 and 1700 and (b) two hundred years of UK Parliamentary Hansard consisting of over 7 million files. To employ a historically valid taxonomy, we have adopted the Historical Thesaurus of English (developed at the University of Glasgow) and the Oxford English Dictionary to help us improve methods for the automatic semantic analysis of historical texts. The Historical Thesaurus contains 793,742 word forms arranged into 225,131 semantic categories. The combination and scale of the corpora and the taxonomy pose significant computational challenges for existing retrieval methods (Wmatrix) and annotation software (USAS) and I will describe our current solutions to these problems.
Theo D. Bothma (University of Pretoria, South Africa): Reflections on the role of corpora and big data in e-lexicography in relation to end user information needs

Corpora are often used by lexicographers to find examples, be this pre-defined corpora or the web as corpus. This is well-known and not discussed in this paper. The use of corpora and big data in digital humanities research also falls outside the scope of this paper. The focus is on how corpora and big data can be used to supplement dictionary data specifically for end user information needs on demand. Research that needs to be done to provide useful tools and the characteristics of such tools will also be addressed. Such supplementary data are not necessarily aimed at the mainstream user of a dictionary who, in a specific work situation, quickly needs the meaning of a word, or a translation equivalent, or who needs to check a specific property of a word (e.g., the gender or a plural form of a noun etc.). It is aimed at the user who is simply interested in a word or phrase, its current and historical usage (especially in context), frequency patterns etc. Within function theory, such information needs are therefore typically not related to text reception or text production information needs in a “standard” work-related environment, but at the level of cognitive information needs, and often simply for edutainment or lexicotainment. However, it could also be useful in specific text production or text reception situations, as will be demonstrated in the paper. Reasons for wanting to access such data will be discussed in brief. Some technical terms (such as big data, structured/unstructured data, text mining, data analytics and the function theory) will be defined briefly and examples will be given.

It is not suggested that any such data form part of dictionary entries, but that links be provided to tools that enable the user to access the data on demand. Two examples of tools will be discussed in this paper, viz. access to linked corpora and access to tools that provide frequency patterns. There are a number of dictionaries that already provide corpus lines on demand, e.g. the DWDS, Danske Ordbogen, BLF, ANW etc. In terms of frequency patterns of word/phrase usage over time, Google has released its Ngram Viewer in 2010. Examples of these products will be given in which the strengths and weaknesses of the current approaches will be discussed at the hand of examples.

To provide detailed and relevant data on demand, a number of technical issues need to be taken into account. The following issues will be addressed: 1) Corpus and data set selection; 2) Corpus clean-up; 3) Corpus markup: corpora need to be marked up at a fairly fine level of granularity. This includes grammatical metadata (e.g. part of speech tagging), as well as bibliographic, genre, date and other metadata. Various aspects of such markup and the reason for the granularity will be discussed. 4) Search functions and filtering of corpus data will also need to be extended considerably, inter alia when spelling variants over time occur. 5) Data presentation and usability issues.

Lexicographers can considerably enhance the user experience by making non-traditional data available to their end users through exploiting the technologies and data accessible through corpora, big data sets and the internet.

Ascensión Gómez Pérez (Universidad Politécnica de Madrid, Spain)
Robert Villa (University of Sheffield, The UK):

Title: Understanding assessors: a forgotten element of big data research?

Open data collections and corpora have been a core part of many different research fields, with annotations of different kinds being generated for a wide range of different reasons. Test collections in Information Retrieval, for example, include assessors’ judgements of document relevance, while there is a wide range of annotated data sets available for use in different machine learning tasks. Reducing annotation effort in building such collections has been widely investigated, with automation and crowd sourcing popular research directions.

When annotations are required, the focus is typically on the generated annotations only: their quality, reliability, etc. Surprisingly, perhaps, there is relatively little research which has considered in more detail the behaviour and background of assessors, e.g. the degree of uncertainty when making a judgement, whether the background of the assessor impacted a judgement, or their behaviour when carrying out the annotation task. While some assessment tasks are relatively objective, many are not, for example it is well known that assessors can often have a widely different view of the relevance of a document to a search topic.

In the project "Understanding the annotation process: annotation for Big data", we have been investigating the human aspect of the assessor in more detail for search topics in Information Retrieval. This has included interviews with assessors and recording assessor behaviour in order to learn more about what different assessors do. From this we aim to develop augmented test collections in which we can embed much more information about the assessors and their judgements beyond the current recording of a binary relevance judgement of “relevant” or “not relevant”. In this talk some initial results from this work will be related to some of the potential advantages of gathering more information about assessors.

Ulrich Heid (University of Hildesheim, Germany): How much does size matter in corpus-based data extraction?

It is well known that statistical methods of NLP (= Natural Language Processing), Machine Learning or Data Mining rely on large amounts of input data and work better on large than on small corpora. For traditional (mainly symbolic) corpus-based annotation, corpus processing or data extraction, the situation is less clear, but seems tendentially to be the same: the more input data, the better the output quality and the richness of the output.

However, for certain languages, domains or types of corpus linguistic applications we either may not have much raw data at all, or we find ourselves flooded with huge amounts of e.g. web texts. Examples of the former situation include specialized languages of particular (e.g. certain technical) domains, historical texts used in Digital Humanities projects, or texts from so-called “under-resourced” languages.

In this talk, we review a number of standard techniques for and applications of corpus annotation, corpus processing and data extraction in terms of their requirements with respect to the size of their input data, as well as in terms of the benefits to be gained from (very) large data sets. We thus briefly discuss tokenizing, pos-tagging and lemmatization, as well as dependency parsing. Our examples from applications concern the identification of syntactic, semantic and collocational properties of lexical items, as well as terminology extraction.
PRESENTATIONS

(1)
Adrover Ginard, Margalida (Universitat de Barcelona, Spain): El catalán septentrional en el Corpus Oral Dialectal

PANEL: CORPUS AND LINGUISTIC VARIATION

En anteriores estudios presentamos parcialmente la variación lingüística del catalán septentrional de finales del siglo pasado a partir sobre todo de los datos relativos a la flexión verbal del Corpus Oral Dialectal (COD) de la Universitat de Barcelona (Viaplana et al. 2007). Perea (2007 y 2009) también trabajó con los datos del COD, centrándose en la descripción de algunos rasgos fonéticos, morfológicos, léxicos y sintácticos del catalán que se habla en la zona francesa del Rosellón.

El COD contiene material fonético y morfológico recogido entre 1994 y 1997 en ochenta y cuatro capitales de comarca o equivalente de los seis principales dialectos del catalán, esto es, catalán central, catalán norte-occidental, catalán septentrional, balear, valenciano y alguerés. La información reunida se obtuvo a partir de cuestionarios de unos seisientos ítems y de un conjunto de grabaciones de habla espontánea de unos diez minutos de duración. Generalmente, se encuestaron tres informantes por localidad, los cuales tenían edades comprendidas entre los treinta y los cuarenta y cinco años y debían a) haber nacido en el lugar; b) pertenecer a una clase social media y c) tener un nivel educativo bajo en cuanto a conocimientos de catalán.

Los datos del catalán septentrional que forman parte del COD se recolectaron en 1997 en cuatro localidades, a saber, Perpiñán, Prada, Ceret y Sallagosa, capitales de las comarcas norcatalanas del Rosellón, Conflent, Vallespir y Alta Cerdaña. Se encuestaron dieciséis informantes.

El presente trabajo pretende complementar los estudios anteriores y analizar la variación lingüística del catalán septentrional de finales del siglo XX a partir de los datos del COD. Concretamente, se describen las siguientes características específicas del dialecto septentrional:

1. Las epéntesis vocálicas y consonánticas finales.
2. La semivocalización de la consonante fricativa prepalatal sorda.
3. La resolución de los grupos gou y gva.
4. El mantenimiento de los grupos románicos N’R L’R y su extensión a casos de NDR etimológico.
5. Las asimilaciones vocálicas y consonánticas.
6. La palatalización de n.
7. La tendencia a la sonorización de las africadas.
8. Los cambios de género respecto de la forma estándar.
9. Las formas del artículo personal y de los pronombres personales.
10. El plural analógico en –os.
11. El paradigma del presente de indicativo del verbo ser.
12. La elisión de la consonante lateral agrupada con ciertas consonantes.
Referencias bibliográficas


PEREA, Maria Pilar (2009): «Elements exolingüés en el català del Rosselló». En: KABATEK, Johannes; PUSCH, Claus D. (ed.): Variació, poliglòssia i estàndard. Processos de convergència i divergència lingüístiques en català, occità i basc.


(2)

Aguiar, Joana (CEHUM- University of Minho, Portugal): A corpus based analysis on causal relations in European Portuguese

PANEL: CORPUS AND LINGUISTIC VARIATION

Although causal connections are well described for Portuguese (Paiva, 1998; Lobo, 2003; Lopes, 2004; Peres & Mascarenhas, 2008; Silvano, 2010), there are not many studies on the frequency of occurrence of these structures in written or in oral texts and on the influence of social variables, such as education level (Lopes, 2004) or gender. Overall, there is still a lack of corpus-based studies on this topic.

Causal relations established between two states of affairs can be asserted or presupposed (Santos-Río, 1976, 1981; Lopes, 2004). Also, presupposed causal relations can be subdivided into explicative (or epistemic) causality and speech act modifier causality (Sweetser, 1990).

Each type of causality can be conveyed by different syntactic structures. Having this in consideration, I propose an analysis of causal relations as a variable phenomenon, which occurrence is constrained by social factors. To verify if and how social factors influence the occurrence of causal relations and their syntactic expressions, two corpora of argumentative texts were gathered and analysed. One corpus is composed by 120 argumentative texts written by request. The other set of texts is composed by 48 texts gathered in blogs.

All texts were written by European Portuguese speakers, stratified according to sex, education (from elementary school to university) and age ([10-12]; [13-15]; [15-19]; [20-45], and [>45]). The themes and the number of words per texts were controlled to avoid biased results.

Each token is encoded according to: type of causal relation (real world causality, explicative causality, and speech act modifier), type of syntactic structure, connector used to establish the causal relation, and position of the causal clause (for adverbial clauses only).

The first finding is that explicative causality is more frequent in texts written by older informants and by informants with more years of schooling. On the contrary, children tend to present causal relations as asserted. Regarding the distribution of syntactic structures, young informants use more adverbial clauses introduced by porque ‘because’. This may
happen because these structures are more common in school contexts (Lopes, 2004) and in question-answer pairs (Diesel, 2004). Moreover, syntactic structures that unequivocally indicate the type of relation established are easier to process (Noordman & Blijzer, 2000). Although there is a general tendency to convey causality through subordination, the percentage of occurrence of causal relations conveyed through juxtaposition increases with the education level. This study also sheds some light on the production of complex structures. It is known that paratactic structures are acquired first (Mithun, 1988; Hopper & Traugott, 2003; Lust et al., 2009). Nonetheless, when used to establish causal relations, juxtaposition and coordination are rare in the texts written by young informants and frequent in the ones written by educated adults. This seems to indicate that the acquisition (and production) of complex structures is also dependent, among other factors (Diesel, 2004), on the semantic nexus conveyed. Regarding the position of the adverbial clause, the work of Mondorf (2002, 2004) shows that women tend to use more causal clauses than men. Also, women seem to postpone the adverbial clauses more often than men do, which may indicate a lower commitment in relation to what is stated (Mondorf, 2002, 2004). Our preliminary results indicate that in European Portuguese there is no significant in the position of the adverbial clause when comparing sentences from texts written by male and female informants.

(3)

Alonso, Isabel & Brouwer, Angela (Universidad Autónoma de Madrid, Spain): Through their own words: A corpus based description of the secondary student teachers’ worries and challenges during their practicum studies at the Madrid Region

PANEL: CORPORAS, LANGUAGE ACQUISITION AND TEACHING

This paper’s main aim is twofold: firstly, to present UAM-ETNA, the corpus of English Teachers’ Narratives, compiled by the DAIC (Discourse Analysis and Intercultural Communication) research group (UAM SOC PR-009) at the University Autónoma of Madrid (UAM) during the last five years. UAM-ETNA intends to fill in the existing dearth of corpora on foreign language student teachers’ language. More specifically, UAM-ETNA provides linguists and discourse analysts with an extensive inventory of lexico-grammatical resources available for the expression of different degrees of professional self-esteem and work satisfaction, basic characteristics of a solid professional teaching identity (Alsop, 2006). And secondly, to illustrate some of the UAM-ETNA potentialities in the context of English as a Second Language (henceforth, ESL) teacher training. To fulfill this second goal, the present paper presents the results of a in-depth discourse study about the main successes, demands and challenges faced by 21 ESL pre-service teachers during their twelve week practicum in different Secondary schools across the region of Madrid. Data is obtained from the analysis of 329 reflective journals (aprox. 90,000 words) which were linguistically annotated and analysed using the UAM Corpus Tool 2.8.12 (O’Donnell, 2012). Findings show that ESL student teachers are mainly concerned with some school issues, namely, how to manage their students’ behavior and how to increase their productivity by choosing the right activities. Less frequently, our prospective teachers reflect on their relationship with other important actors in their practicum studies. Finally, they speak about their personal expectations and feelings concerning their students, the school where they are placed and about the teaching profession in general. Findings are discussed in relation to the most recent research on the development of the new practitioners’ professional teaching identity and on the contextual factors that promote it or hinder it (Beijaard et. al., 2004). On the basis of the results presented, suggestions are made to help set the foundations to establish relevant prevention and action strategies for language teacher education programs, university
supervisors and EFL cooperating teachers.

References

Álvarez Ramos, Eva (Universidad de Valladolid, International Centre for Lexicography, Spain): Uso y desuso del corpus con fines lexicográficos: ¿crónica de una muerte anunciada?

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Cuando en la década de los ochenta la Universidad de Birmingham y la editorial Collins se convirtieron con COBUILD (Collins Birmingham University International Language Database) en pioneras en la introducción del uso de los córpora textuales informatizados (CTI) para la creación de diccionarios, modificaron los parámetros de la metodología lexicográfica: los lexicógrafos podían contar (gracias al corpus) con "evidencias mensurables" (Sindair 1987: XV). La aparición de los CTI fue clave para el desarrollo de la lexicografía, puesto que estas bases de datos ponían a disposición del lexicógrafo informaciones decisivas como las diferentes acepciones de una palabra, las variantes de uso entre la lengua oral y escrita y la frecuencia en el empleo de determinados vocablos (cfr. Pérez Hernandez 2002: s.p.). Su uso fue y ha sido determinante para la creación de diccionarios. Este hecho no significa que ahora el uso de los córpora sea decisivo. Los corpus se crean con el propósito de convertirse en una muestra representativa de una lengua dada, pero por muy rigurosos y amplios que intenten ser, no dejan de ser eso: "una muestra", que nunca llega a representar completa y totalmente la riqueza de la lengua. Ajustar las entradas de un diccionario según se recogen o no en un corpus, lastrea ab nauseam el trabajo del lexicógrafo. En la era de la globalización y de los entornos 3.0, el lenguaje plural, cambiante y lleno matices convierte Internet en una herramienta de trabajo indispensable para el lexicógrafo. La red se erige como el "corpus de los córpora" superándose así las limitaciones del corpus tradicional.

Bibliografía


(5)

Amador-Moreno, Carolina P. (University of Extremadura, Spain): I so like to hear about your little children": exploring the uses of so in the Corpus of Irish English Correspondence

PANEL: CORPUS AND LINGUISTIC VARIATION

Personal correspondence offers a rich and colourful view into the past and has long been used as a valuable source for writing what Elspaß (2005) has called “language history from below” (Sprachgeschichte von unten). In the context of Irish emigration, historians have used personal letters written by merchants, farmers, peasants, artisans and labourers for more than fifty years, but the value of this material for linguistic analysis has not been thoroughly investigated from a variational perspective until recently.

The relative orality of personal letters is also often commented on by both historians and linguists, and recent studies suggest that letters are a possible alternative to the spoken language studied by modern sociolinguists. This paper will use data from the Corpus of Irish English Correspondence (CORIECOR), which contains personal letters from the late seventeenth century to the early twentieth, in order to survey their diachronic development throughout the timespan covered by the corpus (1750-1940). CORIECOR gathers as much evidence as possible for early Irish English into a corpus that permits long-term diachronic study, allowing researchers to trace the emergence and development of features of this variety, including stylistic, regional, and social variation. The corpus can be utilised for empirical comparisons of IrE with data from other sources for the Late Modern English period. A corpus like CORIECOR provides interesting insights into the use of linguistic features generally associated with the spoken mode, such as pragmatic markers, intensifiers, end-tails, etc. The present study analyses the use of so in CORIECOR. It discusses how the use of so is significant in the letters, showing that it was a widespread discoursal feature in Ireland by the 19th century.

(6)

Araújo, Sílvia & Pinto, Paula (Universidade do Minho, Portugal): Fraseología contrastiva, traducción y lingüística de corpus

PANEL: CORPORAS, CONTRASTIVE STUDIES AND TRANSLATION
En el marco de esta comunicación, se trata, en primer lugar, de discutir algunas cuestiones teóricas relacionadas con la investigación en lingüística contrastiva y de poner de relieve el impacto significativo que la lingüística de corpus tiene en la investigación y la enseñanza en áreas relacionadas con el estudio de las lenguas y, en particular, de la traducción. De hecho, nuestra intención es proponer una reflexión sobre la utilidad de un corpus monolingüe y paralelo (Santos & Bick, 2000; Berber Sardinha & Ferreira, 2014) en la práctica de la traducción (Bernardini, 2006; Frankenberg-Garcia, 2009), con base en un ejemplo concreto, lo de las construcciones verbales fijas en la combinación lingüística francés-portugués.

Estas construcciones constituyen un fenómeno muy complejo del punto de vista formal y semántico (Lamiroy, 2008; Mejri, 2008, entre otros) que presenta un cierto interés en el nivel contrastivo, ya que, como trataremos de mostrar, la mayoría de ellas escapan a una traducción directa cuando se buscan sus respectivos equivalentes (fijos o no) en otros idiomas. Es interesante ver que la expresión 'baixar os braços' por ejemplo no da necesariamente lugar a una traducción literal, ya que encontramos, al lado del equivalente más directo de 'baisser les bras', otras formas de traducción más o menos fijas ('jeter l’éponge', 'baisser la garde', 'relâcher ses efforts', ...) que devuelven, cada una a su manera, el sentido (es decir, renunciar a agir) de esta expresión. Una búsqueda en corpus puede proporcionar inmediatamente pistas de traducción mucho más fiables para expresiones cuyo sentido propio no es reconocible por su incongruencia semántica (literalmente, 'couper les cheveux en quatre' es imposible de interpretar; idiomáticamente, esta expresión significa que somos demasiado puntilludos o meticulosos). Para transmitir este sentido figurado, los traductores portugueses tienen que apelar a una amplia red sinónímica, ya que no tienen un equivalente directo en su idioma.

Una búsqueda en corpus ofrece además la ventaja de proporcionar una representación exacta de los diferentes cambios morfosintácticos y semánticos (Mogorrón, 2010) que cada una de estas construcciones verbales puede sufrir en contexto (veamos un ejemplo: 'deitar/meter/lançar/pôr mãos/mão à obra'). Como veremos, no todos los diccionarios incluyen esta variación.

Familiarizar al traductor desde una perspectiva didáctica, con el empleo de herramientas básicas de análisis de corpus (Vargas-Sierra, 2002; Simões, 2008) resulta esencial si se quiere hacer resaltar fenómenos de atracción léxica específicos para cada uno de los idiomas que podrían no resultar evidentes con la única ayuda de un diccionario. Al contribuir al enriquecimiento del diálogo entre dos campos como la fraseología y la traducción, los resultados cuantativos e cualitativos de este tipo de estudio contrastivo basado en (con)textos auténticos pueden dar lugar a la creación de recursos lexicográficos no sólo para la enseñanza de idiomas sino también para los estudios de traducción y, en particular, para la formación de traductores.

Referencias bibliográficas


(7)

**Aurora, Federico (University of Oslo, Norway):** DÂMOS (DATABASE OF MYCENAEAN AT OSLO): annotating a fragmentarily attested language

**PANEL: CORPUS DESIGN, COMPILATION AND TYPES**

This paper presents DÂMOS, the first annotated corpus of all the published Mycenaean Greek texts, allowing for a corpus linguistics approach to the study of the language of the earliest attested Greek dialect (ca. 1450 - 1150 B.C.).

Mycenaean texts are generally administrative documents, written mostly on clay tablets. They have been found within the rests of the Mycenaean palaces both on Crete and mainland Greece. They amount to something less than 6000 documents, but many of them are brief or fragmentary. They are written in Linear B, a syllabic script, not related to the later Greek alphabets, which was first deciphered in 1952, but in scholarly practice they are conventionally transliterated into Latin letters. It is important to remark that although Linear B as a writing system seems to have worked well as a tool for recording and retrieving administrative information, it is not, in fact, a very efficient instrument for rendering the phonetic system of Greek, presenting many inaccuracies and deficiencies in this regard.

The language of the documents, the oldest attested Indo-European language after Hittite and the only attested Greek dialect of the II millennium B.C., presents several archaic and interesting linguistic features and poses some questions crucial for the history of the Greek language (and for the field of comparative Indo-European linguistics), which, especially because of the mentioned limitations of the content of the documents and the shortcomings of the writing system, are still in need of an appropriate, if not definitive, answer.

To create the database, text files with current standard editions as starting point – but extensively revised and updated with new findings, (the numerous) new joins and new readings – have been imported into a relational database (Sql). The texts have then been (partly semi-automatically, partly manually) annotated for morphological, syntactic and lexical information (e.g., the Indo-European root, if reconstructible) for each word, phrase and sentence. A rich set of metadata (hand attribution, find place, chronology, etc.), including detailed epigraphic information on all textual levels (from syllable, to word, line and document level) has also been imported or entered, which is available for searches and can thus be crossed with more strictly linguistic data.
An important feature of DĀMOS is that it allows for multiple analyses of a given linguistic unity to be stored and retrieved. Thus, for example, different hypotheses for the meaning or the grammatical value (e.g. case) of a word can be entered and ranged according to different criteria (e.g. certainty of the interpretation, or scholarly consensus). This feature is, indeed, essential for work with a corpus like the Mycenaean one, where script ambiguities and scantly texts make interpretations often uncertain and dependent on context and intertextual comparison. The linguistic interpretation of a given phenomenon (e.g. the expression of spatial relations in Mycenaean) can, indeed, depend on competing variants of a net of hypotheses (the hypothesized number of cases, the hypothesized phonemic value of certain graphemes, etc.) and implications; it is then crucial to be able to test and compare the different possible linguistic interpretations by varying the value of certain (sets of) analyses in performing complex database queries.

A partial online version of DĀMOS is since February 2013 accessible at: https://www2.hf.uio.no/damos/index/about

(8)
Avila Ledesma, Nancy E. & Romero Trillo, Jesús (Universidad Autónoma de Madrid, Spain): "I am much pleased with this part of the world...": Exploring the Ethnopragmatic Conceptualization of Happiness and Sadness in Irish Emigrants’ Personal Correspondence.

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

The present paper, based on a corpus of Irish emigrants’ personal correspondence, investigates the ethnopragmatic conceptualization of happiness and sadness in the speech of the Irish citizens who emigrated to North America during the eighteenth, nineteenth and the first decades of the twentieth centuries. In particular, this study proposes a Natural Semantic Metalanguage examination (Goddard 2014a) of the emotional load of the positive adjectives (Goddard 2014b, 2014c) happy and glad, and their negative counterparts, unhappy and sad, in order to elucidate the psychological background of Irish migrations to the New World.

The data for the analysis comes from a larger set of emigrants’ letters stored at the Irish Emigration Database and processed using Wordsmith Corpus Tool 6.0. More specifically, it consists of 1153 letters (842,167 word tokens) dating from 1700 to 1920. Following the model proposed by Gladkova and Romero Trillo (2014), the study investigates the most frequent collocations and the pragmatic uses of the adjectives under analysis. For the purpose of this research, the present investigation relies on the semantic explications of happy, unhappy and sad provided by Wierzbicka (1999). Furthermore, we also propose the semantic explication of glad within the theoretical basis of the Natural Semantic Metalanguage (NSM).

In sum, this paper offers a corpus-based study with a new framework for the sociopragmatic analysis of emigrants’ letters. In this regard, the present investigation makes significant contributions to existing research on Irish migrations, being one of the few studies examining the psychological nature of the Irish emigration to North America, from a corpus-based perspective.

(9)
Barrios, María José (Universidad Nebrija, Spain): Análisis contextual en la expresión de incertidumbre: ‘seguro que’ frente a ‘tal vez’. Estudio en un corpus informatizado
El estudio de seguro que y de tal vez que presentamos para este congreso forma parte de una investigación más amplia, cuyo objetivo es el análisis del funcionamiento de los operadores de probabilidad atendiendo a sus propiedades semánticas, sintácticas y pragmáticas.

La expresión lingüística de la probabilidad, fenómeno definible como la manifestación de una incertidumbre orientada, evidencia una gran complejidad en el plano descriptivo, dado el alto número de recursos léxicos – de dispar selección modal – y morfológicos implicados. A fin de efectuar un examen exhaustivo del funcionamiento de elementos tan diversos, se acudió a los textos orales de España del Corpus de Referencia del Español Actual (CREA), donde se analizaron veinte o operadores en 5000 coincidencias. En este estudio analizaremos el funcionamiento de dos operadores distanciados escalonadamente – seguro que, de elevado grado de certeza, en contraste con tal vez, de alta incertidumbre – en un total de 424 textos orales.

La expresión de los operadores de probabilidad desvela su presencia en oraciones del ámbito semántico de la causalidad – concesivas, adversativas, causales, con secutivas y condicionales –, a excepción de las cláusulas finales, donde apenas se detectan coincidencias. Uno de los rasgos más conspicuos del entorno discursivo de los operadores de probabilidad es el acompañamiento de verbos cognitivos (creer, pensar, entre otros), de expresiones dubitativas (no sé, yo qué sé, vete tú a saber, por citar algunos casos) y de distintos operadores en el marco oracional y extraoracional. En menor medida, se encuentran operadores de probabilidad en contextos asertivos. Pese a no tratarse de un fenómeno frecuente, se observa su mayor preponderancia en seguro que, con un 4,6% de casos. Dicha cifra contrasta con el cómputo para tal vez, con un 2,6% de coincidencias. Este fenómeno parece contradecir la esencia misma del operador de probabilidad: ser índice de la falta de compromiso del hablante con el contenido proposicional emitido. No obstante, tal comportamiento obedecería a una intención comunicativa clara: la mitigación de la aserción, conducente a la conservación de la imagen propia y ajena.

Con todo, en relación con lo que acabamos de señalar, una de las propiedades preponderantes de los operadores de probabilidad es su propósito atenuador del acto asertivo, donde tal vez ofrece un 13,2 % de casos, en contraste con seguro que, que cuenta con un 5,1% de coincidencias. En este ámbito, seguro que se caracteriza por presentar casi exclusivamente una usus orientado a prever la imagen positiva del interlocutor, frente a la mayor diversidad de tal vez en este cometido, donde protege tanto la imagen del emisor como la del destinatario.

La revisión de estos operadores desvela una notable asistematicidad en su utilización, concerniente a la escasa influencia de su grado de certidumbre en la mayor o menor ocurrencia de los fenómenos investigados. Con todo, cabe extraer ciertas conclusiones respecto a seguro que en comparación con tal vez. Dicho contraste muestra una escasa coexistencia de marcas de duda junto a seguro que (1,5%), frente a la mayor asiduidad de en tal vez (9,2%), factor que podría deberse al mayor grado de certidumbre del primero frente al segundo. La misma hipótesis daría cuenta de la mayor presencia en tal vez (6,1%) de diversas alternativas a la hora de verbalizar la inseguridad, mostrando distintas hipótesis resolutivas en la explicación de un hecho, frente al 0,5% en seguro que.

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

El estudio de seguro que y de tal vez que presentamos para este congreso forma parte de una investigación más amplia, cuyo objetivo es el análisis del funcionamiento de los operadores de probabilidad atendiendo a sus propiedades semánticas, sintácticas y pragmáticas.

La expresión lingüística de la probabilidad, fenómeno definible como la manifestación de una incertidumbre orientada, evidencia una gran complejidad en el plano descriptivo, dado el alto número de recursos léxicos – de dispar selección modal – y morfológicos implicados. A fin de efectuar un examen exhaustivo del funcionamiento de elementos tan diversos, se acudió a los textos orales de España del Corpus de Referencia del Español Actual (CREA), donde se analizaron veinte o operadores en 5000 coincidencias. En este estudio analizaremos el funcionamiento de dos operadores distanciados escalonadamente – seguro que, de elevado grado de certeza, en contraste con tal vez, de alta incertidumbre – en un total de 424 textos orales.

La observación del comportamiento de los operadores de probabilidad desvela su presencia en oraciones del ámbito semántico de la causalidad – concesivas, adversativas, causales, consecutivas y condicionales –, a excepción de las cláusulas finales, donde apenas se detectan coincidencias. Uno de los rasgos más conspicuos del entorno discursivo de los operadores de probabilidad es el acompañamiento de verbos cognitivos (creer, pensar, entre otros), de expresiones dubitativas (no sé, yo qué sé, vete tú a saber, por citar algunos casos) y de distintos operadores en el marco oracional y extraoracional. En menor medida, se encuentran operadores de probabilidad en contextos asertivos. Pese a no tratarse de un fenómeno frecuente, se observa su mayor preponderancia en seguro que, con un 4,6% de casos. Dicha cifra contrasta con el cómputo para tal vez, con un 2,6% de coincidencias. Este fenómeno parece contradecir la esencia misma del operador de probabilidad: ser índice de la falta de compromiso del hablante con el contenido proposicional emitido. No obstante, tal comportamiento obedecería a una intención comunicativa clara: la mitigación de la aserción, conducente a la conservación de la imagen propia y ajena.

Con todo, en relación con lo que acabamos de señalar, una de las propiedades preponderantes de los operadores de probabilidad es su propósito atenuador del acto asertivo, donde tal vez ofrece un 13,2 % de casos, en contraste con seguro que, que cuenta con un 5,1% de coincidencias. En este ámbito, seguro que se caracteriza por presentar casi exclusivamente un uso orientado a prever la imagen positiva del interlocutor, frente a la mayor diversidad de tal vez en este cometido, donde protege tanto la imagen del emisor como la del destinatario.

La revisión de estos operadores desvela una notable asistematicidad en su utilización, concerniente a la escasa influencia de su grado de certidumbre en la mayor o menor ocurrencia de los fenómenos investigados. Con todo, cabe extraer ciertas conclusiones respecto a seguro que en comparación con tal vez. Dicho contraste muestra una escasa coexistencia de marcas de duda junto a seguro que (1,5%), frente a la mayor asiduidad de en tal vez (9,2%), factor que podría deberse al mayor grado de certidumbre del primero frente al segundo. La misma hipótesis daría cuenta de la mayor presencia en tal vez (6,1%) de diversas alternativas a la hora de verbalizar la inseguridad, mostrando distintas hipótesis resolutivas en la explicación de un hecho, frente al 0,5% en seguro que.
Mostraremos ejemplos concretos de nuestro explicativa del concepto que corresponda al mismo sentido y uso. Esperamos poder ampliarlos y dar así solución a esas entradas.

3. S contextos similares. grupo de palabras que se usen en nuestros corpus para expresar el mismo concepto en lengua francesa en el diccionario multilingüe de turismo COMETVAL.

Las soluciones adoptadas son:

1. Cuando ha sido posible, hemos optado por una equivalencia directa: una palabra o grupo de palabras que se usen en nuestros corpus para expresar el mismo concepto en contextos similares.

2. Sino hemos buscado entre las colocaciones del término.

3. Si hemos hallado escasa o nula presencia de la equivalencia, hemos propuesto el equivalente sin hipervínculo (conscientes de las limitaciones de nuestros corpus, esperamos poder ampliarlos y dar así solución a esas entradas).

4. Cuando no hemos podido hallar la equivalencia, hemos recurrido a una paráfrasis explicativa del concepto que corresponda al mismo sentido y uso.

Mostraremos ejemplos concretos de nuestro sub-corpus en lengua francesa de las cuatro
soluciones adoptadas. Esperamos que esta investigación sirva de ayuda a los usuarios reales o potenciales del diccionario: profesionales del turismo y de la enseñanza de idiomas, traductores y creadores de páginas web, estudiantes y turistas.

Bibliografía:
Fuster-Márquez, Miguel (en prensa) "Lexical bundles and phrase frames in the language of hotel websites." English Text Construction, 7 (1) John Benjamis.

(11)
Bergenholtz, Henning (University of Aarhus, Denmark): A corpus analysis is a superfluous ceremony and a complete waste of your time and the government's money

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The title of this paper does not reflect my own opinion; it is a combination of two quotations from 1976 and 1962. The first part of the title is from a paper from Isa Itkonen, the second one is from Robert Lees. The fight between the position expressed by these quotations and corpus makers was a main topic in the second part of the 1970s. The first book or one of the first books on the construction and analysis of text corpora edited by Henning Bergenholtz and Burkhard Schaefer published in Europe is from 1979, unfortunately with a German title, but with many contributions in English, e.g. by Nelson W. Francis, Stig Johansson, Randolph Quirk and Jan Svartvik. Today, you do not have to fight for making and using corpora. On the contrary, there is a need for discussing the limitations of the use of corpora.

As a lexicographer, I will discuss this problem based on cases such as grammar and collocation items where we definitely cannot get trustful and helpful dictionary articles. But we do not need corpus analysis for all kinds of lexicographic work, e.g. not for the construction of meaning items for some kinds of specialized dictionaries. And we do not need a corpus for the lemma selection in other kinds of specialized dictionaries. But also for general language dictionaries, especially learner’s dictionaries, the need for explicit frequency items is doubtful if the dictionaries are made as tools designed to help the case of text reception or text production.
(12)
Bisiada, Mario (University of Manchester, United Kingdom): Estudio de caso de la metáfora gramatical a través de un análisis corpus de la traducción del inglés al alemán

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

En el estudio de la lengua traducida, a menudo concentramos en el texto final y olvidamos los impactos lingüísticos que tiene el proceso entero de la traducción, especialmente factores extra textuales como los marcados por la dirección editorial. Por eso, fenómenos que observamos en lengua traducida son normalmente atribuidos al traductor, mientras que pocos investigadores (Utka 2004; Munday 2012:110ff; Bisiada 2014) han discutido contrastes entre traducciones publicados y borradores con el objeto de averiguar los cambios que hace el redactor y corrector de estilo. Por ello, esta contribución utiliza un corpus de borradores para destacar la importancia de las “fases de la traducción” (Utka 2004:196).

Eso es demostrado a través de un estudio de caso de la metáfora gramatical ideacional, que, según la definición de Moss y otros (2013:92), “alude a un proceso o cualidad mediante un sustantivo en vez de la realización congruente, que sería, en caso del proceso, un verbo” (véase también Halliday & Martin 1993:141; Steiner 2004), en este caso, en la traducción del inglés al alemán. El foco será en nominalizaciones que son morfológicamente cognados con un verbo de proceso en el sentido de Moss (2013). Se considera el alemán generalmente como una lengua que prefiere un estilo nominal (Nord 1997:60; Fabricius-Hansen 1999:203), el cual logra una densidad de la información más alta que un estilo verbal (Schäffner & Wiesemann 2001:94; Cinto 2009). Sin embargo, algunos estudios han observado que traducciones al alemán, a menudo, realizan las construcciones nominales del texto de partida (TP) como construcciones verbales en el texto de llegada, lo que se ha visto como prueba de explicitación (Konsalova 2007) o “translúcidez” de la lengua de partida, porque “muchas estructuras verbales han sido traducidas literalmente” (Hansen-Schirra 2011:147).

Esta contribución pretende investigar la influencia de los redactores y correctores de estilo en la lengua traducida comprobando la proposición de que hay una tendencia en traductores alemanes a verbalizar estructuras nominales del texto inglés. Para eso, un corpus de traducción, componiéndose (I) de artículos económicos en inglés que aparecieron en la revista Harvard Business Review entre 2006 y 2011 y (II) sus traducciones alemanes (publicados en la revista Harvard Business Manager), será complementado por un corpus de traducciones borradores de estos mismos artículos. Estas traducciones fueron mandadas a los redactores por la agencia de traducciones y por eso representan la lengua traducida antes del proceso de redacción y edición. Éste corpus de 316.000 palabras está alineado de oraciones, que permite una comparación de las construcciones del TP con los borradores y las versiones publicadas. El programa SMOR para el análisis morfológico fue utilizado para descubrir los nominalizaciones alemanes terminando en “-ung” y sus verbos cognados.

El resultado demuestra que los traductores o mantienen el estilo nominal del TP o incluso nominalizan construcciones verbales del TP, como requerirían las convenciones comunicativas del alemán. Crucialmente, observamos que son los redactores y correctores de estilo quienes cambian entonces estas construcciones a construcciones verbales y, por tanto, verbalizan estructuras nominalizadas o restablecen el estilo verbal del TP. Esto permite deducir que algunos fenómenos en la lengua traducida que son discutidos como el “estilo de los traducciones” bien pueden ser atribuidos a los redactores y correctores de estilo. Todo ello, lleva a la siguiente pregunta ¿son adecuados los planteamientos al análisis corpus que sólo consideran la traducción final para analizar la lengua traducida?
(13)
Blas-Arroyo, José Luis (Universitat Jaime I, Spain): El destino de una perífrasis en retirada: la evolución del contexto variable en la selección de haber de + infinitivo entre los siglos XIX y XX. Análisis variacionista de un corpus de inmediatez comunicativa.

PANEL: CORPUS AND LINGUISTIC VARIATION

En el presente estudio se ofrecen los datos de una investigación variacionista para determinar el sino de una perífrasis modal de infinitivo (haber de + infinitivo), arrinconada en la última centuria a contextos cada vez más formales y dialectales en favor de su secular competidora, tener que.

Como parte de un proyecto de investigación actualmente en curso para el análisis diacrónico de las periifrásis modales de infinitivo desde el español clásico hasta la actualidad, para el presente estudio hemos seleccionado un corpus compuesto por textos próximos al polo de la inmediatez comunicativa, escritos por individuos de diferente extracción social y dialectal entre los siglos XIX y XX. Con todo, para facilitar una mayor congruencia de los datos, hemos limitado el análisis a textos redactados por españoles o por individuos nacidos fuera de España, pero que pasaron la mayor parte de su vida en este país. Asimismo, entre los textos se distinguen diferentes registros, que van desde los asuntos más íntimos o familiares en un extremo, a otros de naturaleza menos privada, pasando por diversos grados intermedios. Para el siglo XX, dicho corpus consta de 24 obras, compuestas mayoritariamente por epistolarios privados, aunque también por diversos textos autobiográficos (libros de cuentas, memorias, diarios, etc.). El conjunto, que da voz a más de trescientos cincuenta locutores diferentes, se cifra en 695.090 palabras. Por su parte, los datos del XIX se basan en los materiales proporcionados por 28 documentos del mismo tenor, escritos por un centenar y medio de autores diferentes, lo que representa un volumen total de 490.014 palabras.

En este marco, el análisis exhaustivo del contexto variable que rodea el fenómeno de variación arriba indicado nos ha permitido constatar cómo, pese a los descensos generalizados, y a menudo abruptos, en la selección de haber de con respecto al pasado, existen todavía algunos factores en el sistema que favorecen su empleo, aunque en proporciones mucho menores y siempre por debajo de su competidora. Con todo, más relevante es comprobar cómo, con alguna excepción, la mayoría de estos factores son idénticos a los que han operado en épocas anteriores, aunque esta vez con un descenso en su jerarquía explicativa, así como algunos cambios en la dirección de los efectos, principalmente como resultado del arrinconamiento de haber de + infinitivo en áreas de la gramática y el léxico cada vez más restringidas.

De estos datos se derivan algunas implicaciones teóricas relevantes acerca de las etapas finales de este cambio lingüístico en marcha.

(14)
Botella, Ana (Universitat Politècnica de València, Spain), Gadea, Lucía (Cadena SER/El País, Spain) & Stuart, Keith (Universitat Politècnica de València, Spain): Un corpus periodístico: una metodología para el análisis de la crisis financiera en España

PANEL: SPECIAL USES OF CORPUS LINGUISTICS

En este artículo proponemos un planteamiento metodológico para el estudio lingüístico de un corpus periodístico de la crisis financiera en España. Asimismo, presentamos algunos resultados que están permitiendo elaborar un perfil semántico del sentimiento del periodista. El Corpus de la Crisis Financiera se ha recopilado tomando artículos vinculados
a la crisis financiera del año 2012 en los periódicos de referencia de España - El País y El Mundo - y como representantes en la prensa escrita del bipartidismo; El País relacionado con el partido progresista y El Mundo con el conservador. La investigación que se está desarrollando parte de la idea de que el periodista, en nuestro caso, la prensa escrita, utiliza una serie de recursos léxico-gramaticales en sus artículos, bien para expresar sus propios sentimientos o para ponerlos en boca de los actores principales de sus noticias, esto es, los nombres propios que la crisis financiera ha dejado.

El análisis cuantitativo del Corpus de la Crisis Financiera (CCF), que supone conocer el índice de frecuencias y otros aspectos estadísticos de diferentes elementos contenidos en los textos periodísticos estudiados, nos ha proporcionado datos sobre las palabras y las estructuras utilizadas con un elevado número de apariciones. Además, nos proporciona información sobre los actores principales o entidades relevantes de la crisis financiera, así como el contexto, es decir, el entorno lingüístico en el que dichos protagonistas aparecen (colocaciones y líneas de concordancia). El método de análisis que hemos adoptado nos ha desvelado información sobre determinadas categorías semánticas que co-aparecen con estas entidades: términos de carácter genérico, específico, cuantitativo y cualitativo. Asimismo, se han categorizado diferentes sentimientos, esto es, juicios, emociones y actitudes conforme a los principios de la Teoría de la Valoración (Martin y White, 2005).

No obstante, el análisis por métodos estadísticos ha arrojado algún ejemplo que pone de manifiesto que en los textos el periodista está haciendo uso de estructuras lingüísticas que van más allá de la selección de elementos léxicos individuales, es decir, palabras sueltas. En ocasiones, el autor recurre a unidades semánticas mayores, de una gran riqueza expresiva que requieren un análisis e interpretación más profundos, al margen de las herramientas informáticas utilizadas en las primeras fases del estudio. Por ello, en la siguiente etapa de la investigación nos ocupamos de la interpretación cualitativa con el objetivo de establecer una categorización de sentimientos. A menudo, es evidente que algunos términos son por sí evaluativos, expresando subjetividad y valoración. Sin embargo, en ocasiones, los textos encierran un sentido implícito (por ejemplo, en forma de ironía), enmascarando el carácter evaluativo de ciertas palabras o expresiones. Se han descubierto diversos matices contenidos en el mensaje, que concretamos en una serie de etiquetas que recogen significados que pueden estar presentes en el texto mediante diversas realizaciones lingüísticas o que el lector deberá interpretar o decodificar haciendo uso de su conocimiento previo adquirido. En la producción del artículo periodístico cobra una gran importancia el conocimiento que emisor/receptor comparten sobre el mensaje que se está transmitiendo, en nuestro caso, el conocimiento y la opinión de ambos sobre una misma noticia, que el lector captará de manera inconsciente. Es así como a través de la negociación escritor/lector se forja un sentimiento colectivo que va filtrando en diferentes grupo sociales.

Bibliografía:


(15)

Bouzada Jabois, Carla (Universidade de Vigo, Spain): Tracing the evolution of free adjuncts in English: a diachronic corpus-based description

PANEL: CORPUS AND LINGUISTIC VARIATION

Free adjuncts (FA) are nonfinite (1) or verbless (2) clauses lacking an explicit subject constituent. These constructions are sometimes considered supplements (Huddleston and Pullum et al. 2002: 1250ff) or extra-clausal constituents (Dik 1997: 381) because they are detached from the main clause, they can occupy different positions in clause structure and they are syntactically independent but semantically dependent.

(1) it lays concealed till the prey is entangled, and then coming forth it lays hold on it. (ALBIN-1736,23.646)

(2) A virtuoso in the art of the discourteous aside, he had never been subjected to such disrespect. (Kortmann 1991: 6)

This study aims at providing a complete description of the main features of verbal free adjuncts and analysing their distribution in the Late Modern English (LModE) period. The results will be contrasted with data for Early Modern (EModE) and Contemporary English (PE) taken from previous works by Río-Rey (2002) and Kortmann (1991), respectively.

A corpus-based study has been carried out and examples have been retrieved from the PennParsed Corpus of Modern British English, a syntactically annotated corpus covering data from 1700 till 1914 and divided in three seventy-year periods. I have analysed the first (1700-1769) and the last (1840-1914) periods and have investigated the following features: (i) position with respect to the main clause, (ii) control of the implicit subject by some or any constituent in the main clause, (iii) semantic relation of the FA with the main clause and (iv) presence/absence of introductory elements.

(i) Three positions are available for FAs in clause structure: initial, medial and final.

(ii) FAs have been coded as (un)related depending on the availability of a constituent in the main clause functioning as a proper subject for the FA.

(iii) FAs are said to hold a sometimes unspecified (Thompson 1983: 45, Stump 1985: 1, Kortmann 1991: 1) adverbial relation to the main clause. Two broad semantic categories have been established following Kortmann’s (1991: 121) scale of informativeness: “most informative” vs. “least informative” semantic relations (each of them covering specific meanings).

(iv) FAs are sometimes preceded by introductory elements which are said to restrict the unspecified semantic character of the construction under study. These introductory elements have been categorised in two groups depending on their semantic load.
The data show that most FAs in LModE prefer final placement, and this seems to be in keeping with the data presented in Kortmann (1991: 139) for PE. In what concerns control properties, FAs need to saturate the empty subject slot and, in fact, nearly 90 percent of the examples are controlled by an element in the main clause and more than 85 percent of those related examples show subject-matrix control. Semantically, FAs in LModE show similar proportions for most and least informative meanings; Kortmann (1991: 135) finds a slight increase in most informative examples in PE. Introductory elements in FAs are not the preferred option, with only around 30 percent of the examples in the data showing augmentation.

References:


(16)

Breeze, Ruth (University of Navarra, Spain): Ideology in corporate language: discourse analysis using Wmatrix3

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

Within corpus linguistics, the advent of semantic tagging tools has opened up a new range of possible applications for discourse analysts. Tools such as the USAS function in Wmatrix3 (Rayson, 2008) make it possible to compare the frequency of particular semantic fields in large data sets with appropriate reference corpora, and thus identify the semantic fields that are salient in a particular set of texts. USAS assigns semantic domain tags, which are pre-defined in the underlying lexicon, to the types in a corpus; and Wmatrix3 compares the frequency of the tags thus allocated to particular subsets of the British National Corpus, to ascertain tag keyness via statistical significance (Koller et al., 2008). One of the fascinating possibilities opened up by this is that by examining the key semantic fields identified by USAS for a particular data set, we can detect possible semantic clustering across larger bodies of text, which may enable us to gain a deeper understanding of the way ideology works in specific kinds of discourse.

In this paper, I use the Wmatrix3 USAS tool to detect areas of semantic salience in the genre of the Annual Report. I then draw on the frameworks of discourse analysis and conceptual metaphor theory to interpret the patterns that emerge. Use of this methodology confirmed the expectation that ARs are about “money”, “business” and “numbers”, but also yielded other findings that can be interpreted on an ideological level.
On the one hand, the salience of semantic fields such as “belonging to a group” (teams, networks, etc.) and areas associated with people “in power” (directors, managers, etc.) point to obvious aspects of the ideology of the business world, where the concept of participatory teamwork runs side-by-side with a firm sense of hierarchy (Breeze, 2013). More interestingly, USAS also reveals a considerable clustering around certain value-laden notions such as the idea of “giving” (offer, award, provide, distribute) and “suitability” (relevant, appropriate, qualified), “knowledgeable” (specialist, expertise, recognition, known), “active” (energise, dynamic, interests), “tough-strong” (strong, strengthen, robust), “important” (main, principal, value, significant, priority). These can be interpreted in ideological terms as reflecting utilitarian values, yet framing them positively within a specific metaphorical range (the company is projected as an animate entity that undertakes generous actions, as a powerful physical being, etc.). This is complemented by a high frequency of words denoting “inclusion” (involve, include, integrate, comprehensive) or “helping” (enable, service, boost, aid, benefits), which project a collective identity and an image of the company as a caring family or generous benefactor. Also prominent in the USAS analysis are semantic fields associated with strategic language (plan, strategy, aim, requirement), “investigate-examine-test” (review, research, survey), and “cause-and-effect” (results, consequences, because of, effect, hence, in respect of), all of which shed further light on the utilitarian ideology underlying ARs, and the conventional persuasive framework employed by AR writers to communicate with specific groups of stakeholders. Semantic tagging thus opens up new areas for discourse analysis and the exploration of ideology in text.

(17)

Buda, Jan (Masaryk University, Czech Republic): Collocation of the adjective in Portuguese nominal syntagmata

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

The contribution aims to present a recent research on collocation of adjectives inside nominal syntagmata. The positioning of the adjective has always been subject to many doubts and research on the field of language acquisition as well as normative grammar while, unfortunately, current grammars of Portuguese don’t seem to pay much attention to this problem. The present research examined how its collocation is influenced by diverse linguistic factors, mostly of semantic, syntactic, phonetic or even idiomatic character. The results were confronted with several renowned Portuguese grammars and showed that the semantic factors (most indicated by the grammars) are probably the strongest, but all others are as well present in the decision mechanism of the adjective’s collocation and are mostly ignored by the mentioned grammars.

The research was based on the “Corpus do Português” (www.corpusdoportugues.org) and used statistical methods to prove mainly its hypotheses about idiomatic and phonetic influence on adjectives’ positioning. It also contains a discussion of problems encountered during the corpus processing, which include weaknesses in class definition and distinction (noun / adjective, adjective / quantifier) due to their morphological similarities. A discussion of relevant statistical approach and instruments is included.
(18)

Calvo Cortes, Nuria (Universidad Complutense, Spain): ‘Is he gone’ or ‘has he gone’? On the usage of gone in Jane Austen’s novels and letters

PANEL: CORPUS AND LINGUISTIC VARIATION:

The differences that existed between the use of ‘be’ and ‘have’ as auxiliaries to form the perfect tenses in English, depending on the semantic content of each verb, seem to have disappeared gradually along the Late Modern English period, when ‘have’ became the standard form for all verbs.

The present study focuses on the analysis of the use of either ‘be’ or ‘have’ in combination exclusively with the participle ‘gone’ in Jane Austen’s writings. She has previously been considered conservative in her grammar, specifically in relation to her preference for ‘be’ as opposed to ‘have’ in this type of structures (Ryden & Brorstörm 1987). However, other studies have noted the idiosyncrasy of some of her grammatical structures and have called for deeper analysis to confirm if Jane Austen’s preferences were conservative or not (Tieken-Boon van Ostade 2014).

A corpus-based study of both the novels and personal letters, extracting all the examples of the perfect structures where the participle ‘gone’ is present, was carried out. Whereas the letters showed a higher number of instances of the verb ‘be’ as the preferred auxiliary, the novels seemed to present a more balanced number of examples with either ‘be’ or ‘have’. In an in-depth analysis, it could be observed that the structures in which both auxiliaries are used tend to differ and also that in the novels ‘be’ was the chosen verb when characters were speaking directly or in letters inside such novels, whereas ‘have’ was restricted to the narrator’s voice.

These results led to a two-fold study. On the one hand, the structures were analysed syntactically and semantically. Since the verb ‘gone’ is a clear example of a verb of motion, the main elements taking part in motion situations – Figure, Ground and Path – (Talmy 2000) were analysed and the differences found could most likely explain the choice of one auxiliary or another. On the other hand, the differences observed between the letters and the novels, as well as within the novels, might indicate the manipulation of editors, although they were probably also influenced by the subtleties of the syntactic and semantic differences regarding motion situations.

In addition, the process of grammaticalization experienced by the verbal form ‘gone’ being used later as an adjective most likely contributed to its maintenance in structures such as ‘he’s gone’ in Present Day English.

Finally, a comparison with other writers of the same period would contribute to explaining if her grammatical structures concerning perfect tenses differed or not from the language usage of the time.

Despite the popularity of Jane Austen’s novels, she has not received a lot of linguistic attention so far (Tieken-Boon van Ostade 2014). This study will, therefore, provide a better understanding of her use of the language both in her more private writings and in her novels, and it will also explain the linguistic reasons for the particular choice of one auxiliary verb or another in different cases.

References

The objective of this paper is to carry out a comparable corpus-based analysis of the common linguistic features of English to Spanish translations and investigate the potential linguistic and textual constraints of using computer-assisted translation systems.
The initial hypothesis for this empirical work lies in the assumption that the choice of the segmentation rules of most translation memory systems affects the translator’s approach to the target text and it is at the editing stage where these sentence shifts need to be rectified and adapted to a more natural target language characteristics.

The first part of this work is devoted to the revision of the literature on translation strategies in order to identify those strategies which require further processing in terms of syntax and change in sentence structure. Secondly, we identify the characteristics of computer assisted translation systems and contrast them with the selected translation strategies that require further editing in the target language and therefore more relevant to working with CAT tools; within this part, we concentrates on providing a quick overview to the translation workflow with CAT tools and the description of their most characteristic functions, especially those related to the interaction and effects with the translator’s linguistic decision-making processes, such as segmentation rules.

The third part of the work exemplifies the different findings with a bilingual comparable English-Spanish corpus.

Finally, the last stage emphasizes the synergies between the findings during the literature review on translation strategies and identifies translation strategies potentially affected by the use of CAT tools.

The methodology includes the description of the compilation process of an aligned bilingual corpus. After aligning the segments, the next step consists in identifying the equivalences and the solutions proposed by the translator.

Among the conclusions, it should be pointed out that it is at the editing stage, which is given increasingly more relevance due to the momentum and quality improvement of machine translation, when the translator/ editor has the possibility to correct and rectify translation decisions made during the work with CAT tools. Finally, this paper presents a typology of the strategies most commonly used during the CAT tool translation process.

(21)

Cantos, Pascual & Almela, Moisés (University of Murcia, Spain): The interaction of collocational and colligational priming in corpus-based descriptions of conceptual networks

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Previous research into lexical constellations has uncovered the existence of dependency relations among different collocations of a word (Cantos & Sánchez, 2001; Almela, 2011; Almela et al., 2011; Almela, 2014). Such dependencies obtain when the strength of the attraction between a node and one or more of its collocates is contingent on their co-occurrence with a third element (a co-collocate). For instance, the probability that face (v.) collocates with decision is increased by the presence of modifiers of a specific semantic type (e.g., hard, difficult, tough) but weakened by the presence of other types of modifiers such as wise, informed, rational, etc. Implications of this phenomenon for the analysis of word meaning and for the notion of ‘collocation’ have been examined in previous studied.

With this paper our goal is to explore an aspect of co-collocation that has not been tackled yet, namely, the relationship with colligational priming. In Hoey’s (2006) theory of lexical priming, colligational priming refers to the grammatical profile of a word’s combinatory behaviour. Our study is motivated by the hypothesis that patterns of co-collocation will exhibit regularities also at the grammatical level and that, if this is so, the dependency relations observed between collocates and co-collocates will also be reflected in
dependency relations found at a more abstract level of analysis, i.e., between different grammatical slots in the environment of the node. Our analysis will be focused on the three most common syntactic types of collocation: subject-verb, object-verb, and modifier-noun. Using the nouns cause and basis as nodes, we will first determine whether they tend to occur more frequently in subject or in object position, and if so, whether the observed tendency is correlated with a tendency to occur in collocations where they modify other nouns or in collocations where they are modified by other nouns or by adjectives. The lexical association score used is logDice (an improved version of Dice); the scores will help us determine the inter-collocability dependency; that is, given Prob1=Prob(collocate1|node, collocate2) and Prob2=Prob(collocate1|node); if Prob1>Prob2 then we find positive inter-collocability; else negative inter-collocability (Prob1<Prob2). All the data used in this research have been extracted from the corpus enTenTen [2013], accessed at SketchEngine.

References


(22)

Casas-Pedrosa, Antonio Vicente (University of Jaen, Spain): Differences between spoken and written English: the case of the predicative Prepositional Phrases in the ICE-GB

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

This paper is aimed at describing the main differences between spoken and written English. More specifically, attention is paid to the different examples which are classified as predicative Prepositional Phrases (PPs) in the International Corpus of English-Great Britain (ICE-GB) and their frequency in spoken and written texts. These units can be defined as those phrases which are introduced by a preposition and followed by a Noun Phrase (NP) acting as its complement. Furthermore, they perform the function of Subject Complement (Cs) at clause level. Such is the case of “She first fell in love with Will when she was eighteen, and she adores him still” (ICE-GB:W2F-019 #47:1).
Although in terms of frequency this is not the syntactic function PPs more often perform, they are taken into account because of their complexity and due to the lack of detailed analyses. In most cases they are described as isolated examples and this phenomenon is not considered to be a very productive one.

After introducing some basic notions, these structures are analyzed focusing on their presence in both spoken and oral texts within the ICE-GB. This is a one-million-word corpus which is both morphologically tagged and syntactically parsed. Moreover, it was compiled in the nineties and consists of both spoken (60%) and written material (40%).

The ICECUP (ICE Corpus Utility Program) software retrieved 3307 examples from 3223 sentences. These instances were then filtered since some of them were later classified as “noise” (in some cases the PPs were performing other functions either at phrase or at clause level and in others the element acting as the complement of the preposition was not a NP). For these reasons the final subcorpus consists of 1332 examples.

67.49% of these instances (899) are found in oral texts whereas 32.51% of them (433) belong to written texts. All these examples have been classified into different groups and subgroups corresponding to the different text categories available in this corpus (Nelson, Wallis and Aarts, 2002: 307-8). The results are presented in charts by means of both figures and percentages and different conclusions are later drawn based on the analysis of these charts.

Thus, for example, it can be noticed that, although it was expected that the amount of structures under study would be higher in spoken than in written texts because of the structure of the corpus itself, the relative frequency (which takes into account the relationship between the number of examples and the number of words) proves so, too: 0.1410% in spoken texts as opposed to 0.1022% in written texts, with an average of 0.1255% in the whole corpus. Moreover, there are more examples in dialogues (581) than in monologues (318) and in printed texts (332) than in non-printed ones (101).

This information proves especially relevant for non-native speakers of English since it allows them to become aware of the differences between speaking and writing. According to the evidence, some units are used more often in spoken language than in written English. Therefore, when producing any kind of text, students will feel more confident for they will be able to choose the appropriate structures bearing in mind these issues.

(23)

Casas-Pedrosa, Antonio Vicente (Universidad de Jaén, Spain): The economy principle and English predicative prepositional phrases

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

This paper is aimed at analysing the relationship between the economy principle and English Predicative Prepositional Phrases (henceforth, PPPs). These can be defined as those phrases which are headed by a preposition whose Complement (C) is a Noun Phrase (NP) and which perform the function of Subject Complement (Cs). Such is the case of “under arrest” in [1]:

[1] The vessel remained under arrest from September the twenty-sixth until October the nineteenth (ICE-GB:S2A-065 #18:1:A).

As for the economy principle and the principle of the least effort, Vicentini (2003) studied the origin of these theoretical notions. Different examples obtained from the BNC and ICE-GB corpora and from various dictionaries confirm the hypothesis according to which the selection of certain PPPs will allow speakers to convey a given meaning by means of a
more reduced number of words. Thus, the PPPs “in clover” and “in hand” are defined as “to have enough money to be able to live a very comfortable life” (Turnbull, 2010: 279) and “receiving attention and being dealt with” (Crowther, 1995: 537), respectively:

[2] “As I was saying,” Patrick Milligan continued, once his youngest was out of the house, “if the best came to the best, and your sister married the old codger, we could be in clover” (BNC:EEW 2057).

[3] In fact the repairs were already in hand <n> (ICE-GB:S1B-069 #163:1:B).

These sentences clearly show that PPPs which are formally simple can express complex ideas. In fact, “in clover” and “in hand” illustrate the smallest structure of a PP, just consisting of a preposition and a NP as its C.

However, on some occasions certain PPPs are chosen to avoid redundant structures such as “be being”. In fact, the use of “at issue” and “under construction” in [4] and [5] prevents speakers from saying “may be being dealt with” and “which will be being built”, respectively:

[4] Again, the meaning of `necessary` may be at issue but the important factor is that the presumption can be and, in many cases, probably will be cancelled out by express terms (BNC:HDX 175).

[5] One of the major features is a timber-framed house which will be under construction throughout the show, allowing visitors to see the various elements and skills involved (BNC:A16 61).

Furthermore, there are other reasons why PPPs are selected in certain communicative contexts. Thus, “in the club” is defined as “pregnant” (Rundell, 2007: 273; Turnbull, 2010: 279) and “off your chump” as “crazy” (Rundell, 2007: 240), but these PPPs also convey some other subtle nuances. For that reason they are labelled as “British informal old-fashioned” (Rundell, 2007: 273 and 240, respectively). Therefore, it can be concluded that these are counter-examples since in some cases speakers will opt for more complex structures (“pregnant” and “crazy” are one-word adjectives, whereas the PPPs “in the club” and “off your chump”, on the contrary, consist of 3).

To this last group of examples belong some PPPs which can be classified as euphemisms. Rees (2006: v) defined them as follows: “[...] the word or phrase has the specific function of concealing something of the nature and meaning of what it describes”. Such is the case of the PPP “in Abraham’s bosom” in [6], which could be replaced by the adjective “dead”:

(24)

Choudhary, Prakash (National Institute of Technology Manipur, India), Nain, Neeta & Ahmed, Mushtaq (Malaviya National Institute of Technology Jaipur, India): A Linguistic Structure to Develop and Annotate Urdu Corpus for Multidisciplinary Research on Urdu Handwritten Documents

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

In this proposal, we are describing a methodology for building an Urdu corpus CALAM (Cursive and Language Adaptive Methodology) including a large volume of 1200 handwritten text image file. A language independent structure has been designed to annotate handwritten Urdu script image for higher level at lines, words, components level with a XML standard to provide a ground-truth of each image at different four levels of annotation in a standard encoding UNICODE UTF-8. For capturing maximum variations in Urdu words and balanced the corpus, data collection is distributed within 6 categories and
further divided into 14 subcategories and forms were filled by different writers from various geographical regions with different educational qualifications. The structure mapping provides facilitates of corpus creation and navigation through all the information of handwritten images and segmented lines, words and components very easily through structure without losing the broad view of the input data and provide additional support for: Insertion, modification, Classification of corpus data and searching for direct access of the needed attributes and annotated information.

Over the past few years’ lot of advancement have been made in the field of handwritten recognition techniques, linguistic resources such as annotated corpus are playing a significant role and are the most demanding platform for computational linguistic research. Computer processable corpus linguistic has more capability to explore and identify all features of natural language including the characteristics of the desired texts such as lexical, textual, semantic and syntactic attributes.

The generation of corpus methodology for Indian languages has started in 1991, but compared to other languages very less attention is given to Urdu language. A standard database is essential part for learning and evaluation of various OCR techniques and in automatic data entry for a complex script like Urdu where text entry requires more effort as compared to English, as an input of a single character in Urdu text requires a combination of multiple keys. Urdu language is historically related with India from the time of Mughal Empire. It is the national language of Pakistan, and one of the 22 scheduled languages in the Constitution of India. India is a country with a large number of native Urdu speakers in its five states: Andhra Pradesh, Jammu and Kashmir, Bihar, Uttar Pradesh and New Delhi. The population of Hindi-Urdu speakers is the fourth largest community in the world after Mandarin, Chinese, English and Spanish. According to Govt. of India 2001 census data [Census 2001] around more than 50 million persons speak Urdu in India. Urdu is the official language of Jammu and Kashmir State and recently Urdu is also approved as the second official language of Uttar Pradesh.

In this work we are presenting a large Urdu handwritten text corpus database having full length Urdu sentences and annotation structure which has been experimented for Urdu, Hindi and English for annotating offline handwritten image corpus with four level XML presentation. The annotation of corpus is essential to make it available and applicable in a vast area of computational linguistic. Here, a unified approach is used to develop an Urdu corpus along with the demographic information of writer on a single form. Urdu is the fourth most frequently used language in the world but due to its complex writing script and poor resources it is still a thrust area for NLP. We have developed CALAM (cursive and language adaptive methodology) an Urdu corpus consisting of 1200 handwritten images. For capturing maximum Urdu words and the variations in handwritten styles data collection is distributed within 6 categories (History, Literature, Science, News, Architecture and Politics) and further divided into 14 subcategories and forms were filled by different writers from various geographical regions with different educational qualifications. A structure has been designed to annotate handwritten Urdu script image at lines, words, components level with a XML standard to provide a ground-truth of each image at different four levels of annotation in a standard encoding UNICODE UTF-8 which is highly desirable as shown in Figure 1.
Figure 1: Hierarchical process of corpus design and annotation

The structure mapping provides facilitates of corpus creation and navigation through all the information of handwritten images and segmented lines, words and components very easily through structure without losing the broad view of the input data and provide additional support for: Insertion, modification, and searching for direct access of the needed attributes and annotated information.

The aim of the structure is to build a resources that would provide ground truth annotation for multilingual handwritten text images. Structure would be rich source to design a large volume database for natural language processing related research. This corpus would be very useful to provide researchers all the facilities for linguistic research on same platform such as: benchmarking and evaluation of handwritten text recognition techniques for Urdu script, signature verification, writer identification, digital forensics, and classification of printed and handwritten text, language teaching, and categorization of texts as per use and so on.

Clavel Arroita, Begoña & Gregori-Signes, Carmen (IULMA, Universitat de València, Spain): Analysing lexical density and lexical diversity in university students’ written discourse

PANEL: CORPORAS, LANGUAGE ACQUISITION AND TEACHING

The terms lexical density, lexical diversity or lexical richness (Daller et al. 2003) refer to lexical measurements employed to estimate the lexical wealth of texts and thus can also be used to assess students’ progress regarding the acquisition of vocabulary. It is generally assumed that texts with higher density are harder to understand and that spoken texts have lower density levels than written texts (Ure, 1971; Halliday, 1985). Nevertheless, a text may show high lexical diversity in other words, it may contain many different word types, and low density at the same time if it contains many pronouns and auxiliaries as opposed to nouns and lexical verbs (Johansson, 2008).

Researchers in SLA have always been interested in the analysis of students’ written production as part of their overall communicative competence. A more precise knowledge of lexical density and diversity, when obtained through reliable quantitative and qualitative analyses can provide teachers and researchers with a clearer picture of students’ progress. Lexical richness can be a factor to be taken into account when designing teaching materials and assessing students. In our view, these principles justify the analysis described here, which is part of a larger longitudinal study in the framework of a research project taking place at the Universitat de València (CASTLE- GV/2014/022). The CASTLE project aims at studying the lexical development of students over several years, with the purpose of uncovering possible patterns in their progress.

The subjects in the study are first year students enrolled in a Preliminary English Test (PET) course, that is, B1 level according to the CEFR, in the academic year 2012-2013. The first composition was collected during the second week while the second task was collected six weeks later. The subjects in Corpus C are students who are taking a course in C1-C2.

Our study takes into account both lexical density and lexical diversity in three corpora: a) Corpus A. Essay1, by A2-B1 students; b) Corpus B. Essay2, by A2-B1 students; c) Corpus C. Essay by C1-C2 students. Each composition comprising around 200 words, a total of 135 compositions. Corpus C is used as the reference corpus.
The main goal of the study was to compare the lexical density and lexical richness at the beginning and end of the teaching period in order to assess their CEFR level at those two stages. Lexical density was tested using Textalyzer (http://textalyzer.net). The tool employed to test lexical frequency was the software RANGE (Nation and Heatly, 1994). Our results prove that both corpora show the same progression between writing 1 and 2. Furthermore, we can claim that it is possible to obtain a reliable measure of lexical richness which is stable across two pieces of writing produced by the same learners.

Finally, thanks to the use of RANGE and Textalyzer, we could measure the Lexical Frequency Profile (Laufer and Nation, 1995). The LFP shows the percentage of words a learner uses at different vocabulary frequency levels, in other words, the relative proportion of words from different frequency levels. We found that the LFP correlates well with an independent measure of vocabulary size. This reliable and valid measurement of lexical richness in writing can be useful for determining the factors that affect judgments of quality in writing and for examining how vocabulary growth is related to vocabulary use.

References


(26)

Clua, Esteve (Universitat Pompeu Fabra, Spain) & Lloret, Maria-Rosa (Universitat de Barcelona, Spain): El COD2: un corpus oral para el análisis de la variación espacial y temporal del catalán

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

En esta comunicación presentaremos el COD2 (Corpus Oral Dialectal del Catalán Contemporáneo 2), que es la continuación del Corpus Oral Dialectal del Catalán Contemporáneo de la Universidad de Barcelona, y que nos permitirá realizar un análisis de la distancia lingüística de las variedades de la lengua catalana tanto des del punto de vista espacial como temporal.

El COD2 forma parte del proyecto ADLET (Análisis de la distancia lingüística en los ejes espacial y temporal: aspectos fonológicos y morfológicos del catalán), que a su vez es la continuación del proyecto DiLET, en cuyo marco se recogieron los datos orales del corpus.

El objetivo del proyecto es contribuir a ampliar el conocimiento sobre la variación lingüística en general y sobre la distancia entre variedades lingüísticas, en particular,
from a perspective double: the spatial and the temporal. This is part of the
limitation of the linguistic distance among the Catalan varieties. For this reason
we are constructing the COD2, as the more updated version of the Catalan dialects.
This corpus will serve as a base, in the first place, to define the linguistic
existence in the present among the geographical varieties of the whole
linguistic area, by means of an automated dialectometric analysis in which we
will use different techniques of quantitative analysis: the MCD (COD Method),
the VDM (Visual Dialectometry), the DIATECH, etc.; and, secondly,
with the COD, we can determine the evolution of this distance and, in that
sense, the linguistic experiment conducted.

Although only two decades have passed since the establishment of the COD,
we consider that during this period the school education in Catalan and the generalization of
the standard language in audiovisual media, at least in some regions of the
linguistic area, have allowed to provide processes of leveling and dialectal
convergence that have been studied. We propose, therefore, starting from
the data collected in the previous project, to constitute a corpus that will be
representative of the Catalan varieties, which is comparable with the COD,
and bring to the dialectometric analysis of the data to determine the linguistic
distance spatial and temporal of the varieties of Catalan.

(27)

Colwell, Verónica (Universidad de León, Spain): Meaning-making and
language learning resources for real-life purposes: corpus-based writing applications

PANEL: CORPORAL, LANGUAGE ACQUISITION AND TEACHING

Despite the recent remarkable headway made here in Spain where the
learning, teaching and assessment of English as a foreign language (EFL)
is concerned, today many small and medium sized businesses still struggle to
tackle the challenges posed by English as a lingua franca in the marketplace.
Whether they need to sell face-to-face or virtually in the global
village if they wish to survive at all in this ever-more competitive world,
optimal communication skills and proficiency in English, particularly in writing,
remain one of the most important pressing challenges such companies face.
Hence, the need arises for reliable applications and tools that support those
foreign language users in the workplace by meeting their specific requirements.
This paper reports on a case study, carried out
with a small group of L1 Spanish undergraduates in the EFL class beyond B1,
which set out to explore and illustrate some of the ways recently designed
corpus-based writing applications, developed by the project group XXXX to assist professionals in the workplace
in their endeavours to produce top-quality texts in English (Rabadán 2005-2008), serve
not only to assist lifelong language learners engaged in tasks in real-life learning
environments but also to extend and develop the language knowledge and skills of EFL
learners in formal classroom settings, as it is precisely within these settings that meaning-
making tasks which emphasize language learning for real-life purposes have an
increasingly more important role to play (Swales 1990). Indeed, the genre-bound nature
of the writing applications which draw on corpora specially built for the production of
effective and practical tools for specific work environments, certainly go a long way
towards meeting the needs of language learners in formal settings in a number of ways,
not least by providing an ideal platform for efficient and reliable consciousness-raising
and language awareness-raising activities (Ellis 2006; Sinclair 2004; Swain & Lapkin
1995; Wichmann et al 1997). Our results indicate that the corpus-based writing
applications described in this report not only help to bridge gaps in the linguistic
repertoire of intermediate and advanced language learners but furthermore that they
represent a rich language learning and meaning-making resource for engaging such
learners in relevant, real-life tasks.

30


(28)
Conejero-Magro, Luis Javier (Universidad de Extremadura, Spain): La intertextualidad bíblica de los dramas históricos de Shakespeare a la luz de la lingüística de corpus

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

Este estudio es una aproximación al sentido de la intertextualidad bíblica en los dramas shakesperianos. En concreto, el análisis se centra en las alusiones y las referencias oblicuas a la Biblia. Los datos manejados proceden de la obra de Shakespeare y serán cotejados con un corpus de referencia de carácter bíblico. La utilidad de las técnicas y los instrumentos de la lingüística de corpus aplicados a los datos procedentes de esos corpus y a los aportados por la crítica sobre el discurso bíblico en Shakespeare ha sido doble. En efecto, por una parte han permitido corroborar el origen auténticamente escriturístico de la totalidad de las frases y las unidades fraseológicas contrastadas, mostrando infinidad de ejemplos idénticos o parecidos en otras obras de Shakespeare cuyo contexto bíblico puede resultar más aclaratorio. Por otra, el resultado de este análisis reafirma la tesis de que ese uso frecuente de alusiones o referencias al Viejo y al Nuevo Testamento en ningún caso tiene en la obra de Shakespeare una finalidad de tipo doctrinal o de naturaleza ética sino únicamente estética o estilística. De un análisis contrastivo posterior entre esos focos de intertextualidad bíblica del texto isabelino y los de las traducciones españolas de José María Valverde y Luis Astrana Marín se desprende que estos autores, y de manera más sobresaliente José María Valverde, logran recrear ese intertexto bíblico que tanto contribuye a la configuración de los personajes principales en las obras de partida.

(29)
Corino, Elisa (University of Turin, Italy) & Panunzi, Alessandro (University of Florence, Italy): Corpus-based collocational patterns in different domains: a tool for lexicography and LSP
The paper will deal with corpus-based extraction of VERB-NOUN (V-N) collocational patterns in different domains, with the purpose to show how the same collocation can be differently used and can assume different shades of meaning according to the context where it is occurs.

The VERB-NOUN collocation seems to be more binding than NOUN-ADJECTIVE or VERB-ADVERB chunks and it is the most likely source of errors among both natives and learners, due to the synonymic drift, causing inappropriate combination of words which can thwart comprehension.

Even if strict collocations are included in general monolingual or bilingual dictionaries - or they are organized in collocation dictionaries, it often happens that the domains where they are used and the register they pertain to are not mentioned in the dictionary article.

Even in collocation dictionaries the differences between strict V-N collocation and free collocational pattern are not clear-cut and little is said about their pragmatic uses. Tiberii (2012), for instance, puts strict and free collocations together in the same section, whereas both Urzi (2009) and Lo Cascio (2013) organize them according to their syntagmatic structures adding some brief context, though none of them provides the domain –and the register - where the collocation is mainly used.

The original RIDIRE corpus (Panunzi, Cresti&Gregori 2014) serves this purpose. The contribution will presents possible uses in the fields of lexicography, winking at the potential employment in language acquisition and language learning.

The RIDIRE corpus is a balanced Italian web corpus (1.5 billion tokens) designed for both lexicographic purposes and enhancing the study of Italian as a second language, chiefly as far as collocational patterns are concerned. The targeted crawling was performed through content selection, metadata assignment, and validation procedures. These features allowed the construction of a large corpus with a specific design, covering a variety of language usage domains (News, Business, Administrationand Legislation, Literature, Fiction, Design, Cookery, Sport, Tourism, Religion, Fine Arts, Cinema, Music). The query system allows research to be carried out on the whole corpus itself or on the sub-corpora. Specifically, available queries comprehend all the functions usually exploited in corpus-based lexicography: frequency lists, concordances and patterns, collocations, Sketches, and Sketch Differences.

The research will be carried out among highly polysemic words, such as the Italian noun campo, which, as well as its English counterpart "field", can refer to a variety of referents: e.g. a "farmed field", a "sport ground", or the "field of vision". This polysemy is reflected by the collocational behaviour of this word within different linguistic domains. Using RIDIRE search tools, we can compare the Sketches of the word campo in the domain of Sport and Cinema (and it could also be extended to News and Business). Among the most prominent verbs occurring before this noun in the Sport domain, there are espugnare, sbaragliare, violare, sbancare, all verbs that allude to an overwhelming victory in an away game (also by means of military metaphors). On the contrary, in Cinema domain we find verbs like restringere, allargare, sgomberare, ingombrare, which are related to the concept of "framing" in cinematography.

REFERENCES


RIDIRE Corpus Online. Accessed at: http://www.ridire.it

(30)

Creese, Sharon (Coventry University, United Kingdom): In Search of Oblivion? How the ‘Right to be Forgotten’ Could Undermine Web-Based Corpora

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

This paper discusses the potential difficulties facing builders of web-based corpora as a result of rules1 surrounding the ‘right to be forgotten’ or ‘right to oblivion’, recently upheld in a case against internet search engine, Google (De Terwangne, 2012: 111; Cox 2014).

Linguistics researchers know well the need to approach search engine numbers with caution, since their focus is on generating maximum hits, regardless of how tangential they are to the original search term (see for example Fletcher, 2013: 3). Since May 2014, however, researchers have faced the added complication of potentially useful sources being excluded from results, because of the ‘right to be forgotten’.

The ‘right to be forgotten’ means that if a story is deemed inaccurate, if the information is no longer relevant, or if it is considered excessive, individuals have the right to request that links to it be removed by search engines (Travis & Arthur, 2014).

Between May and September 2014, Google received some 130,000 such requests, many involving cases of fraud, violent crime and child sex abuse. Approximately 50% of these were upheld without challenge, 30% were investigated and 20% rejected (Cox 2014).

Questions are already being raised about the appropriateness of these new rules – including alleged inconsistencies in rulings – however the implications for researchers building web-based corpora are even more significant (Ibid). The data collection phase of my PhD study, entitled ‘Exploring New Words in Online Newspapers and in Wiktionary: Lexicographic and Linguistic Perspectives’, demonstrated the kinds of difficulties involved.

My work involves building a corpus of newspaper articles appearing in the online versions of five UK national newspapers:
Google Advanced Search (GAS) was chosen to locate them, however problems arose when searching The Sun.

Several of the 36 neologisms returned null results through GAS, despite the newspapers’ own search engine having located articles containing them. These contradictory results raised questions about the possibility of missing links, leading to the supposition that the links had actually been removed. Yet not all links for all words were missing. Whilst none of the results found in The Sun for ‘floordrobe’ or ‘diabetesy’ were included in GAS, seven of the 36 results it returned for ‘frenemy’ were. At the same time, several of the articles did not appear to fit the criteria for removal. This raised the question of whether some links might have been taken down in error; whilst this cannot be proved, no explanation other than the ‘right to be forgotten’ could be found for their absence.

In the case of my research, the issue was resolved by removal of The Sun from the list of newspapers. To check if any of the others were affected, the number of results returned by GAS was compared with those from internal search engines; consistency across the two indicated they were not.

Such a simple solution will not always be available however. As more and more requests for ‘oblivion’ are approved, and the danger of accidental removal grows, the number of texts available for corpus research could be eroded. Over time, this could lead to an increasingly ‘patchy’ picture of language use, akin to a digital image marred by missing pixels.

Endnotes
1. Part of EU Directive 95/46 on Processing Personal Data
2. www.thesun.co.uk

Bibliography

(31)

Cristobalena, Araceli (University of León, Spain): A corpus-based genre study of instruction manuals for household appliances

PANEL: SPECIAL USES OF CORPUS LINGUISTICS

Whenever we open a household appliance (or any other mechanical or electrical product, such as a toy) we can find a user’s manual; this can be a leaflet or an 80-page book. In any case, we agree to name it instruction/user’s manual and we treat it as a unit. This communication presents a research done with the purpose of establishing the rhetorical structure of instruction manuals for household appliances and to contrast the differences in two languages, Spanish and English. With this commitment we have compiled a 52-text bilingual corpus divided into two comparable subcorpora following Sinclair (1991) and Leech (1996) ideas on corpus linguistics as well as other scholars.

The theoretical frame for this contrastive study is based on macro-linguistics. We have taken the notion of genre to approach the corpus and considered instruction manuals as one more of them. This macro-linguistic perspective is based on Swales (1993), Halliday & Hasan’s (1989), Werlich (1983) and Bhatia’s (1993, 2004) notions. Then, as the communicative process (Shannon, 1948) and the pragmatic relations are similar in every case, we have been able to preform a rhetoric analysis from a twofold perspective: qualitative and quantitative.

On the qualitative side of the analysis, we have determined the prototypical structure, following and adapting Swales model (1993) to the instructional genre, and according to it we have tagged the corpus into moves, steps and substeps. On the quantitative side of the analysis, we have looked for the significance of the different moves, steps and substeps in numerical terms and percentages. This raw information needs interpretation, and the conclusion will extract it from the differences and similarities in terms of priorities in one language and the other.

References


(32)

Diez Bedmar, María Belén (Universidad de Jaén, Spain): Exploring the use of the English article system in Literature papers: Spanish and American senior undergraduates compared

PANEL: CORPORAS, LANGUAGE ACQUISITION AND TEACHING

The research which analyses the use of the article system by ESL or EFL students has increased since the publications in which the article system was considered a grammatical morpheme (Hakuta, 1976; Huebner, 1979; 1983; Tarone, 1985). Later, the presence of articles in obligatory contexts was discussed, although without referring to a detailed subclassification of the contexts in which those articles were employed. It was in the 1980s when this gap in the literature was bridged with Bickerton’s (1981) use of the binary semantic and discourse-pragmatic features, that is, speaker reference [+SR] and hearer’s knowledge [+HK], together with Huebner’s (1983) subsequent taxonomy. These two publications resulted in the division of the article use into four types of contexts, namely type 1 contexts [-SR, +HK], type 2 contexts [+SR, +HK], type 3 contexts [+SR, -HK], and type 4 contexts [-SR, -HK].

Since Bickerton’s and Huebner’s publications, many studies have used their taxonomy to analyse the students’ use of the article system (see, for instance, Ekiert, 2005; Humphrey, 2007; Diez-Bedmar and Papp, 2008; Diez-Bedmar, 2010; Haiyan and Lianrui, 2010; Nickalls, 2011; Diez-Bedmar and Pérez-Paredes, 2012; Diez-Bedmar, 2015; etc.). In the case of Spanish EFL learners, the research which has used this taxonomy has reported findings concerning a) an Integrated Contrastive Model (ICM) (Granger, 1996; Gilquin, 2000/2001) with Spanish and Chinese students (Diez-Bedmar and Papp, 2008); b) the comparison of article use by secondary school leavers and first-year university students (Diez-Bedmar, 2010); and c) a cross-sectional analysis of the article use by secondary school students at all levels in compulsory and optional secondary education (Diez-Bedmar and Pérez-Paredes, 2012); and d) a cross-sectional analysis regarding CEFR A2 to CEFR B2 levels to find out possible criterial features (Diez-Bedmar, 2015).

Although these results are important in the description of the Spanish students’ profiles and their language learning acquisition process, there have been no studies so far analysing Spanish university students’ use of the article system in academic writing, more specifically, Literature papers. Due to the increased input received by Spanish senior undergraduates, and the vital role played by academic writing in the students’ careers, two main objectives were pursued in this paper. First, to conduct a CIA (Granger, 1996) with the written production by American and Spanish senior undergraduate students to analyse the type of contexts and articles employed by both learner groups; and b) to conduct an IL analysis (Selinker, 1972) to explore the use of the article system by the Spanish learner group and find out the problems that the use of the article system in the FL poses at this advanced level.

To do so, two learner corpora were employed. First, a subsection of the longitudinal learner corpus compiled and error-tagged in Jaen (Spain), which comprises literature half-term exams (33 texts, 45,876 words) by fourth year university students. Second, a
subsection of the *Michigan Corpus of Upper-Level Student Papers* (2009), containing Literature texts written by fourth-year university students (26 texts, 46,333 words), was used as control corpus. The analysis of the article system was conducted with the tagging system developed by Díez-Bedmar and Papp (2008).

The results obtained allow a) the analysis of the way how these two learner groups conceptualize and express the same ideas in different ways; and b) insights into the problems found in article use by Spanish university students concerning the articles used and the contexts in which they are employed when writing their Literature exams before finishing their BA.

References:


(33)

Dotti, Fiorella Carla (Universidad Autónoma, Madrid): Project Arcturus: A tool to teach historical linguistics

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

Some knowledge of historical linguistics is expected of students of linguistics. This area presents a big challenge to most students of linguistics because of either a lack of sufficient practice or a lack of a solid foundation in phonology and phonetics. In order to alleviate this problem, a teaching tool was developed to assist in the practice of phonological and orthographic evolution since Indo-European until Present Day English.

Previous tools were aimed mostly at comparison, such as LingPy (List et al, 2014) or Iberochange (Eastlack, 1997), and many are not functional today, such as PHONO (Hartman 1993). Previous approaches geared specifically at teaching this subject were not found.

The goal is not only to help in the acquisition of knowledge in historical linguistics, but also to foster scientific thought and to serve as a first approach towards scientific linguistic research.

With this in mind, the tool can be used in two different ways: teachers can preset the rules that they want students to practise, or students can be instructed to deduce and create their own rules, and then simulate the evolution. The rules are defined in terms of patterns, such as whether changes took place in all vowels or only long vowels or in a student-defined group of characters/phonemes, in what period the change occurred, whether shifts were involved, whether there were exceptions, the phonetic/orthographic context in which the rule applies, etc. This forces students to identify the underlying linguistic principles in order to create a rule. Since rules are stored in separate files for each student, and many rulesets can be used by one student, it allows for theory development and testing without affecting the work of other students. Sharing rules is possible if this setting is enabled by the teacher, in order to promote teamwork. Rulesets can be downloaded in a non-proprietary format, so that researchers and students can back them up and view them outside the application as well.

When carrying out evolutions, each step is corrected automatically, to prevent students from dragging an error throughout the whole evolution. If provided with an appropriate corpus, students can see the word in context as it evolves. They could write the Old English form of the word and the program can identify the corresponding Middle English or Present Day English forms, for instance, and provide an example in the form of a keyboard in context. This allows students to track semantic changes through history.
In order to restrict the challenge to the subject matter and not extend it towards configuration and encoding issues, the program is a web application that can be accessed from an internet browser and it allows the user to input IPA symbols in Unicode from an online keyboard.

The tool is currently being improved towards dialect identification, and it can interact with large corpora of English dialectal writing. The program also makes use of statistical techniques to identify and suggest the most prominent features of English dialects, if provided with a sufficient number of texts corresponding to that dialect.


(34)

Eurrutia Cavero, Mercedes (Universidad de Murcia, Spain): Huella cultural en el léxico sobre inmigración: estudio comparativo entre los subcorpora normativo e informativo de los documentos jurídico-administrativos emitidos por las Administraciones Públicas francesas

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

El presente estudio se enmarca en el Proyecto de investigación interuniversitario y multidisciplinar concedido por el Ministerio de Economía y Competitividad sobre El lenguaje jurídico y administrativo en el ámbito de la extranjería: estudio multilingüe e implicaciones culturales (LADEX) y tiene por objeto contribuir a rellenar el vacío existente hasta el momento, sobre trabajos de investigación científico-filosóficos, en lengua francesa, que conjuguen la caracterización del lenguaje jurídico administrativo con las implicaciones sociales y culturales de las migraciones internacionales. Con el fin de satisfacer dicha demanda lingüística y social, profundizaremos, en los aspectos sociolingüísticos e implicaciones culturales derivadas del uso restrictivo del léxico y de la fraseología en los textos propios de este ámbito, emitidos por las Administraciones Públicas francesas desde 2007 hasta la actualidad.

La metodología de investigación seguida se basa en la lingüística de corpus. Tras la compilación y sistematización de diferentes corpora de documentos de diversas tipologías discursivas, de uso obligado para todo ciudadano extranjero, disponibles en formato electrónico, procederemos a su clasificación taxonómica según se trate de textos emitidos por la administración y dirigidos al ciudadano (FR-1), actos que la administración destina a la propia administración (FR-2), textos informativos (FR-3) o actos realizados por el ciudadano y destinados a la administración (FR-4). Posteriormente, seleccionaremos dos subcorpora para el estudio contrastivo que proponemos basándonos especialmente en su relevancia en el sector específico abordado. En nuestro caso, nos apoyaremos en los subcorpora constituidos por los textos normativos (LADEX-FR-1) e informativos (LADEX-FR-3) centrándose en el análisis de los nombres abstractos.

El procesamiento de los archivos mencionados a través de la herramienta Sketch Engine permitirá examinar la frecuencia de uso normalizada del lenguaje que aparece en los mismos y facilitará la selección de los términos de mayor frecuencia de uso y cuya carga

Para la anotación subsiguiente proponemos proceder en primer lugar, a la categorización del nodo respecto a sus colocaciones atendiendo al tipo de asunto administrativo en cuestión (entrada, legalización, residencia, coste económico, amenaza, regreso...), factores culturales, aspectos religiosos, culturales, de entidad étnica o nacional; cuestiones referentes a atributos diferenciadores como edad, sexo, familia, trabajo; marcas específicas en alusión a un determinado grupo u organización. En segundo lugar, nos centraremos en la categorización gramatical del nodo en su contexto discursivo respondiendo a anotaciones sobre su función de sujeto, de modificador, si aparece coordinado con otro término... Por último, analizaremos la prosodia haciendo especial énfasis en las connotaciones positivas, negativas u otras, asociadas al contexto en el que se actualiza el lema.

Concluiremos pues cómo la caracterización comparada del usuario extranjero e implicaciones culturales refleja sin duda el estado de la cuestión en los ámbitos lingüísticos detallados y contribuye a definir la condición del extranjero en la sociedad francesa incentivando el debate sobre la solidaridad, desde la perspectiva del lenguaje.

(35)

Farquharson, Joseph T. & Galarza Ballester, Teresa (University of Bielefeld, Germany and TBA, Spain): The Adequacy of Newspaper Corpora for the Lexicographical Description of New Englishes

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Over the past two decades there has been a significant increase in the popularity of corpus methods as a means of achieving descriptive adequacy on the basis of empirically sound methods. However, this revolution has done very little to improve the plight of undocumented and under-documented languages and varieties. In fact, most existing corpora are for large standardised languages (e.g. English, Portuguese, Spanish) which already have a long tradition of descriptive and prescriptive grammars. The languages and varieties that would stand to benefit the most from a corpus linguistics approach tend to be those whose speakers lack the financial, material, and human resources to apply the approach in an effective manner.

Caribbean varieties of English fall into the category of the severely un(der)documented. This probably stems from the fact that for a long while after gaining political independence from Britain, many English-official Caribbean countries considered themselves users of [Standard] British English. Therefore, there was no need to study the variety of English used locally. Findings from the International Corpus of English (ICE), sub-corpora for Jamaica, Trinidad and Tobago, and The Bahamas reveal that this view is not supported by the available evidence. The ICE corpora are generally too small though to allow for the volume of data that a good lexicographic description requires. However, online versions of newspapers provide a stand-in measure since they offer the lexicographer a searchable corpus that is often free of cost.
This paper explores the adequacy of using online newspapers as corpora for the lexicographic description of Jamaican English and Antiguan English, two varieties of so-called new Englishes spoken in the Caribbean. It explores how much ground can be covered by using a single mode (written) and a single genre (news writing) in working out the use and definition of a selection of words. The newspapers to be used are the Jamaica Gleaner for Jamaica and the Daily Observer for Antigua. The words to be surveyed include basic, mostly Anglo-Saxon words such as god, mother, mouth, kill, eat, heavy, heaven. For the first part of the research the newspaper corpora will be searched for the selected words and the various senses will be worked out based on the available data. The second part of the research will compare the range of meanings found from the corpus work to the range of meanings encountered in the Oxford English Dictionary (OED) and the Merriam-Webster Online Dictionary (MWOD). Additionally, a survey will be conducted among users of Jamaican and Antiguan English to ascertain how many of the meanings listed in the OED and the MWOD for the selected words are used by or known to them. This will give us a standard of comparison to determine how effective the single mode/genre corpus is in the work of a lexicographer. This discussion will also highlight some of the pitfalls to be considered when conducting such work.

(36)
Favaro, Alessandro (Università degli Studi di Padova, Italy): The Particular Negative: a Distributional Study on Some Aspects of Meaning Contradicting Logical Equivalence

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

The main aim of this paper is to reflect upon some aspects of meaning related to two different ways of expressing the same proposition type in English. The proposition type under discussion is the particular negative, and it is accounted for as introduced by either 'not all' or 'some' followed by a verbal negation. Two statements serving as examples of the two expressions are:

1- Not all birds can fly;
2- Some birds cannot fly.

The theoretical assumption underlying this paper is that two logically equivalent quantified expressions might not be equivalent from certain semantic and conversational points of view, and that, conversely, two expressions that are equivalent from a conversational point of view might not be equivalent from a logical one. The research has been carried out by means of a distributional study: by exploring the contexts of occurrence of our two particular negative expressions throughout the ukWaC corpus we have detected some systematic use and aspects of meaning specific to each of them.

Before tackling the data, the paper provides a detailed account of how the distributional analysis has conducted out and of the aspects which have been focused on. The analysis consisted of a thorough exploration of almost five-hundred contexts of occurrence of 'not all can' and 'some cannot', and the main aspects we focused on are related to:

1- the quantity with which the two expressions (and their alternatives) are most likely to be associated;
2- which of the two expressions is more likely to be specified by means of examples within its context of occurrence;
3- which one is more frequently used within disjunctive structures;
4- which one is more likely to be based on presuppositions related to the same predicate
as applied to 'all'. (For example, for statements such as 'not all humans are honest' or 'some humans are not honest' to be understood properly, they need to be related to the presupposition that 'all humans are honest' represents an ideal situation.)

Next, the data yielded by the research are stated and discussed. Some of the most significant observations following the analysis are:

1- 'not all can' is followed by a reference to its complement (i.e. to those who/which can) more frequently than is 'some cannot' (12.4% of the contexts of occurrence for 'not all can' and 6% for 'some cannot'); moreover, such alternatives are very often associated to quantifiers such as 'most', 'many', 'the majority of';

2- 'some cannot' is more frequently followed by examples specifying who or what the quantifier 'some' refers to (14.9% of the contexts for 'not all can' and 6.3% for 'some cannot');

3- both 'not all can' and 'some cannot' are very often used within disjunctive structures (34.1% and 21.3% of the contexts);

4- presuppositions related to 'all'(of any of the three kinds we considered) underlie more frequently the use of 'not all can' than that of 'some cannot' (8.1%, 9.4%, 6.8% vs 2.5%, 2.3%, 4.3%).

The very last part of the paper is a brief account of the words occurring most frequently within the sentences including either 'not all can' or 'some cannot': the fact that the word 'example' proved to be one of the thirty words occurring most frequently with the second expression seems to be further evidence that 'some cannot' is more likely to be accompanied by explicit examples.

(37)

**Faya Ornia, Goretti (Spain):** El folleto médico traducido al inglés en España. Estudio contrastivo con corpus.

**PANEL:** CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

El objetivo de esta ponencia es dar a conocer los resultados obtenidos en un análisis contrastivo del género del folleto médico. En primer lugar, hemos extraído y comparado las características de los folletos originales en español y en inglés. En segundo lugar, contrastamos estas características con las que presentan los folletos traducidos en inglés en España. De este modo, podrremos determinar si los folletos traducidos a inglés en España están influidos por los folletos originales españoles o si siguen los rasgos propios de la cultura meta. Para extraer las características de cada grupo de documentos, hemos trabajado con un corpus de 300 documentos (100 folletos en español, 100 en inglés y 100 traducidos).

(38)

**Fernández-Domínguez, Jesús (IULMA, Universitat de València, Spain):** Determining changes in availability over time

**PANEL:** CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The term available is used in word-formation studies in reference to those morphological processes which can be employed to create new words at a given time. Availability is often
depicted together with profitability (the number of lexemes that a process can coin), and together they embody the hyperonym productivity. Availability is in this context portrayed as the qualitative side of productivity, with profitability representing its quantitative side. This explicit division of productivity into availability and profitability was not made until Corbin (1987), and the former still stands as a slippery concept.

Unambiguously as the above may seem, the definition of availability is in itself challenging for various reasons. One of them is its widespread denominative overlap with the label productive, which makes it often difficult to discern different uses of the term. Nonetheless, the notion of availability gets particularly problematic when we turn to diachronic word-formation because the fact that a morphological process proves to be available at a given time in the history of a language does not ensure that it will be available at a later stage. This involves not only that “statements of availability are temporally limited” (Bauer 2001: 205; see Bauer et al. 2013: 32), but also that there must exist a locus for the transition from availability to unavailability (and vice versa). If, as is widely agreed, availability is a non-gradable concept, research on word-formation should be able to detect that hinge and describe its nature. This need is perceived as growing if we consider accounts like Bauer et al.’s (2013: 198-201), who report on morphological processes that allegedly stopped being productive in the past but where lexical activity seems to have revived (for example, -ment or -th).

With the above in mind, this paper studies changes in the availability/unavailability divide from Early Modern English by resorting to the OED and to diachronic corpus data. In the specialized literature different methods have been tested to measure productivity over time and expounded their pros and cons (Plag 1999, Cowie & Dalton-Puffer 2002, Palmer 2014, Säily 2014). In this case, the morphological activity of various processes is pinned down in the OED and subsequently tracked in the Helsinki Corpus, which allows detecting similarities and differences between the two methods. Morphological productivity is here explored in a number of time frames starting in Early Modern English, thus making it possible to observe changes in a process’s historical behavior.

This analysis evidences, among other things, that the figures obtained for low productivity differ depending on the perspective and methods adopted for the investigation. Likewise, the comparison of dictionary and corpus data allows describing and contrasting the fluctuations in the efficiency of processes whose productive status has been described as uncertain, for example, derivation by -ment (Bauer et al. 2013: 198-201, Palmer 2014). This study lastly facilitates an assessment of the creative power of such processes and questions how the transition from and to unavailability may be empirically determined.

(39)

Fernández-Fuertes, Raquel, Álvarez de la Fuente, Esther (University of Valladolid, Spain) & Liceras, Juana (University of Ottawa, Canada): The acquisition of sentential subjects: an analysis of English and Spanish monolingual and bilingual corpora

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

Bilingual child language acquisition research has long been concerned with whether interlinguistic influence between the two languages of the bilingual shapes the non-adult patterns of omission/production of functional categories (Müller 1998; Döpke 2000; Yip & Mathews 2000; Hulk & Müller 2000; Paradis 2001; Nicoladis 2002; Paradis & Navarro 2003; Zwanziger et al. 2005; Serratrice et al. 2009; Fernández Fuertes & Liceras 2010; Liceras et al. 2012, among others). In this paper, we analyze the omission/production of subject pronouns in the developing English grammar and in the developing Spanish grammar of two English-Spanish simultaneous bilingual children, one English monolingual
child and one Spanish monolingual child. We base this analysis on (i) the divide between two different reformulations of the null subject parameter (Alexiadou & Anagnostopoulou 1998 and Sheehan 2006); and (ii) Liceras et al.’s (2012) assumptions concerning the role of “lexical specialization” in interlinguistic influence.

Children acquiring English and Spanish produce cases of subject omission, as in (1) and (2), in spite of the fact that null subjects are possible in adult Spanish but not in adult English.

1 (it) Roars [Simon, 2;05 (FerFuLice corpus)] English = [-null subject]

2 Tengo más [Manuela, 1;11 (Deuchar corpus)] Spanish = [+null subject]

When comparing the monolingual and bilingual acquisition of English and Spanish the questions that arise are the following: (i) will monolingual and bilingual children show similar patterns of production/omission of subjects in English?; (ii) will monolingual and bilingual children show similar patterns of production/omission of subjects in Spanish? If monolinguals and bilinguals pattern differently, (iii) can this be accounted for in terms of interlinguistic influence? And in that case, (iv) which language is the locus of interlinguistic influence? If Spanish null subjects influence English, then bilingual English would display more null subjects and for a longer period of time than monolingual English; if English overt subjects influence Spanish, then bilingual Spanish would display more overt subjects and for a longer period of time than monolingual Spanish.

To provide a comparative analysis of the patterns of production/omission of English and Spanish subjects, three longitudinal corpora available in the CHILDES project (MacWhinney 2000) have been selected: the FerFuLice corpus which corresponds to the spontaneous production of two bilingual children (Simon and Leo); the Sachs corpus on data from one monolingual English child (Naomi); and the Ornat corpus on the production of one monolingual Spanish child (María). Given that the study focuses on the analysis of the earliest stage (onset) English grammar and of the earliest stage (onset) Spanish grammar, the age range covered is from 1;06 to 2;11 years.

The results show that monolinguals and bilinguals exhibit similar patterns of subject omission/production in Spanish but that these patterns are, however, different in the case of English. In particular, and with respect to interlinguistic influence, while English does not influence the overproduction of overt subjects in Spanish, the presence of null subjects in Spanish has a positive influence in the eradication of non-adult null subjects in bilingual English (as in 1 above). We argue that in a bilingual situation, as compared to a monolingual one, lexical specialization in one of the languages of the bilinguals (i.e. the availability of an overt and a null realization of the subject to satisfy the requirements of Spanish) facilitates the acquisition of the other language.

(40)
Fernández de Molina Ortés, Elena (Universida de Burgos, Spain): El mantenimiento de las variedades dialectales en las redes sociales

PANEL: SPECIAL USES OF CORPORA

La presencia de las redes sociales en la sociedad actual ha modificado por completo el uso de la lengua escrita; la comunicación oral ha traspasado el papel y se ha implantado en la interacción comunicativa en la red. En este trabajo se utilizará un corpus compuesto por comentarios obtenidos de las redes sociales Facebook y Twitter entre los años 2011 y 2013 para analizar cómo los usuarios de estas plataformas de comunicación hacen uso de
sus variedades diatopicas tanto en su muro como en las conversaciones directas e indirectas que mantienen con otros usuarios de la red social. Además, para observar la consciencia lingüística de los hablantes en la escritura en las redes sociales, se presentarán los datos obtenidos de las encuestas realizadas a los usuarios de estas plataformas de comunicación.

(41)
Fuyuno, Miharu, Yamashita, Yuko & Nakajima, Yoshitaka (Kyushu University, Japan): Head movements and speech pause insertion patterns in English public speaking performances: Investigation on a multimodal corpus data of Asian EFL learners

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

Reflecting the expansion of the field of corpus linguistics, not only text-based written language data but also other dynamic features of language communication such as phonological aspects and gestures have started being analyzed over the last two decades (Knight, 2011; Adolphs & Carter, 2013). Corpora with these multi-channel input are generally referred to as multimodal corpora.

The present study is a part of our longitudinal multimodal corpus project being conducted at Kyushu University. The project has been aiming to construct and analyze a multimodal corpus of English public speaking performance by Asian EFL learners. The database has been compiled with audio- and video-recorded data of English recitation performances presented in an official recitation contest held at a Japanese university. In addition to the performance data, evaluation data by official contest judges are also stored in the database, enabling us to examine performances in consideration of their effectiveness.

In our previous research on this project, Fuyuno et al. (2014a; 2014b) analyzed multimodal data of six EFL learners from the corpus to find out that speakers who had received higher scores from contest judges had similar pause distributions as native speakers of English, whereas speakers with lower scores did not.

Based on these previous results, this present study analyzes performances of other six EFL learners from the corpus. The previous studies focused on speech pause duration, speech rate and head movements of the speakers. In the present analysis, the pause positions in relation to grammatical chunks are also examined as well as the pause length and head gesture factors.

Methods of acoustic and movement analysis were used to examine the data. Regarding the audio data, acoustic analysis software (Praat) was used for automatic extractions of pause positions and lengths from each performance. To analyze the video data, 2D computer-vision based motion tracking was conducted to obtain the head movement patterns of the speakers.

By analyzing the data, the following results were identified; (1) Speakers who received lower evaluation from judges shared similar pause-distribution patterns, especially regarding pause insertions around connective phrases. (2) Speech rates and pause distribution patterns indicated a correlation. (3) Tendencies of head gestures in regard to performance evaluation scores are similar to the results of the previous studies. The results of the present study are expected to be fruitful to improve the practical teaching of English-public-speaking teaching for Asian students.


(42)
Giacomini, Laura (University of Heidelberg, Germany): Building Sample LSP Corpora for Translation

PANEL: CORPORAS, CONTRASTIVE STUDIES AND TRANSLATION

In performing translation tasks involving texts with a moderate degree of specialisation, such as newspaper texts on specialised topics, translators primarily rely on LGP dictionaries, LSP dictionaries and online documentation. The general lack of publicly available specialised corpora and of appropriate LSP lexicographic products confronts professional translators as well as translation students with the issue of retrieving specialised knowledge from heterogeneous and often outdated sources. At the same time, online documentation, primarily concerned with the identification of parallel texts, often represents the only efficient means of finding satisfactory contextual equivalents. The paper aims to illustrate the results of research tasks related to corpus methods in translation, in which different working groups tested the suitability of small-sized, sample corpora as additional, ad-hoc reference resources.

(43)
Giacomini, Laura (University of Heidelberg, Germany): Context-dependent variation of LSP collocations: A corpus-based analysis

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

This paper illustrates how a specific sub-LSP corpus can be employed to study and explain collocational variation, a phenomenon that is largely underrepresented in bilingual LSP lexicographic resources. Terminological variation on the formal and semantic level is looked at from the point of view of the relation between concept and designation. This is useful especially for less standardised subdomains, like many technical subfields, in which text production is often hampered by the presence of several domain synonyms with no easily predictable contextual distinctions. A balanced monolingual LSP corpus makes it possible to evaluate variation as a context-dependent phenomenon, with context interpreted as the combination of topic, genre-specific and communicative functions.

(44)
Lafuente Giménez, Sabrina (Universidad Jaime I, Spain): Corpus oral e investigación de la interferencia en el discurso de docentes brasileños de español como lengua extranjera: Estudio de caso
PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

Sobre las características de las lenguas próximas, algunos estudios empíricos recientes (Durão e Andrade, 2010) estiman que el español y el portugués comparten aproximadamente un 85% del léxico en varios campos semánticos. Otros autores como Almeida Filho (1995), afirman que entre las lenguas románicas el portugués y el español son las que tienen más afinidad. En consecuencia, en razón de la similitud (morfológica, sintáctica, semántica y fonética fonológica), no existen alunmos luso hablantes considerados como “principiantes” en español (Carmolinga, 1997), visto que normalmente ya tienen adquirida la capacidad de comprender parte del idioma, tanto en el registro hablado como en el escrito. Por otro lado, y, paradójicamente, una de las mayores dificultades de dichos estudiantes es superar las similitudes existentes en los dos idiomas, que terminan por facilitar las interferencias de la lengua materna en la extranjera.

No siempre es fácil, aún en niveles muy altos de conocimiento de la lengua extranjera, desprenderse de algunos aspectos propios de la lengua materna. En el caso de los docentes de español hay que añadir la preocupación por la influencia que las interferencias puedan ejercer en el aprendizaje del alumno. Por esa razón, este estudio pretende averiguar si en el discurso oral de la población analizada -profesores brasileños de español como lengua extranjera- se presentan signos de interferencia de la lengua materna. El trabajo se apoya en la lingüística de corpus, a través de la cual se establecen los límites temporales, geográficos y lingüísticos del corpus estudiado (Atkins, Clear, & Ostler, 1992). La colecta de datos se hace por medio de la grabación y transcripción de clases y entrevistas en audio que conforman el corpus oral (Du Bois, 1991). Se delimitarán los análisis en el marco de las subcompetencias gramatical y léxico-semántica y se usará el método de análisis de errores apoyado también en la lingüística contrastiva sincrónica. Su carácter es cuantitativo y cualitativo, buscando, por un lado, estudiar la frecuencia de los errores y, por otro, describir y detallar los diferentes tipos (Corder, 1981) con base en la gramática funcional.

BIBLIOGRAFÍA


(45)

Giménez Folqués, David (Universidad de Valencia, Spain): Anglicismos originales y sus adaptaciones en corpus del discurso turístico español 2.0
La difusión del turismo en español, como en otros idiomas, ha ido en aumento gracias a su internacionalización y a su expansión en Internet. Tanto es así, que los corpus turísticos han aumentado su caudal léxico, y, por lo tanto, han proporcionado al discurso turístico una tendencia creciente en cuanto a interés lingüístico. Este discurso turístico está repleto de extranjismos, principalmente de anglicismos, ya que, debido a su internacionalización, ha encontrado en estas voces una manera de expresarse y llegar al máximo número posible de usuarios. Estos extranjismos suelen aparecer en su vertiente original, aunque, debido a las características de la lengua española, podemos encontrar algunos de ellos adaptados en grafía y pronunciación.

Para llevar a cabo este trabajo, vamos a partir de corpus turísticos elaborados en el proyecto Análisis léxico y discursivo de corpus paralelos y comparables (español-inglés-francés) de páginas electrónicas de promoción turística, desarrollado en el grupo de investigación COMETVAL, de la Universidad de Valencia. A partir de este material, vamos a intentar dar luz sobre la situación de los anglicismos en el discurso turístico español. Estos corpus comprenden el ámbito panhispánico, en el cual, además, se ha pasado el filtro de las recientes obras académicas (el Diccionario panhispánico de dudas y la Ortografía de la lengua española) para conocer si responden a un patrón normativo.

En resumen, sacaremos conclusiones sobre el uso de anglicismos en el discurso turístico español 2.0. acerca de:

- la razón de su uso;
- si se usan adaptaciones o su forma original;
- y si se usa la forma normativa recogida por los diccionarios académicos o no.

(46)

Godinho Soares Vieira, Nataliya (NOVA University of Lisbon, Portugal): E-learning practices in interpretation and translation: corpora as training platforms

PANEL: CORPORAS, CONTRASTIVE STUDIES AND TRANSLATION

To study another language to be professional translator or interpreter is a challenging mission. It demands good practices in a wide range of sectors. The linguistic structures, cultural markers, the limits of linguistic resources for expressing ideas – all these elements require the continuing improvement. Following the discussions in Corpus-based translation studies [Bowker, 2000; Laviosa, 2002; Zanettin et al, 2003; Corpas Pastor, 2008; Kruger, 2011; Vargas-Sierra, 2012; Gallego Hernández, 2012; Vieira G. S., 2014], this paper is designed for the purpose to highlight the multifunctional competence of corpora to assist autonomous trainings in interpretation and translation. The paper focuses on the following aspects: 1) classification of some written or spoken problematic segments for transmitting from one language to another that can be easily acquired through autonomous corpus-based trainings; 2) selection of audio and visual concordances for simulating simultaneous interpretations; 3) observation of parallel concordances for identifying the most relevant translation equivalents. For this reason, the paper explores some monolingual, parallel and audiovisual corpora as the English Language Interview Corpus as a Second-Language Application (ELISA), the Michigan Corpus of Academic Spoken English, the Scottish Corpus of Texts and Speech (SCOTS), Russian Multimedia Corpus, Parallel corpus of English and Portuguese (COMPARA), Corpus Paralelo CLUVI, Corpus del Español, The POLYU Language Bank, etc.
(47)


PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Modern-day lexicographic practice is characterised by its reliance on speech corpora as significant and non-negotiable sources of support to obtain material for inclusion in dictionaries. This applies to both printed dictionaries and e-dictionaries. Corpora are utilised for a variety of lexicographic purposes, e.g. to compile frequency lists from which the lemma candidates can be selected, to find orthographic variants of the lemma candidates, to find the necessary morphological guidance like plural and diminutive forms and to supply illustrative material that can be included as items giving cotextual guidance.

Items from which cotextual information can be retrieved fall into different categories with example sentences, the typical citations from corpus material, as one type, indicating the macro- syntactic environment of the word represented by the lemma sign, and collocations, indicating its micro- syntactic environment. Both these types of text segments are usually included in the same search zone, i.e. the article slot for illustrative examples.

Too often dictionaries do not make a distinction between these types of cotextual items and the user has no clear guidance when looking for a collocation. For text production purposes, for both mother-tongue and foreign speakers of a given language, collocations have to be seen as a necessary items and they should be presented in a clearly identifiable way. Their lexicographic significance may never be underestimated.

Collocations need to be presented as part of the cotextual guidance of a given lemma. However, they should not only be entered as items addressed at the lemma sign but should rather be elevated to the level of secondary guiding elements, functioning as treatment units in their own right and as an address of additional microstructural entries.

A separate search zone should be introduced to accommodate collocations and their treatment. This treatment should consist at least of a cotextual occurrence of the collocation and, where necessary, a collocation of collocations. The single collocations sincere desire and express a desire can collocate to form the complex collocation to express a sincere desire. This kind of guidance can really benefit users with a text production need and lexicographers could utilise corpora not only to find a collocation of the word represented by the lemma sign but also to find data on the typical use of such a collocation. In a bilingual dictionary an equivalent for the collocation could be given and the corpus of the target language could help to determine whether this equivalent is also a target language collocation or merely an item giving a non-collocational translation.

This paper gives a discussion of the way in which corpora can be used in this regard. Corpus evidence in the selection and treatment of collocations can play a determining role in the realisation of the need for dictionaries to respond to the real needs of real users. It is also shown how this realisation can lead to an expansion of the article structure in order to accommodate an additional search zone dedicated to collocations.

(48)

Gris Roca, Joaquín (Universidad de Murcia, Spain): Categorización de actividades usadas en la Enseñanza de Inglés como Lengua Extranjera a partir de un corpus de materiales didácticos

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING
El objetivo de este estudio es establecer una tipología de las actividades docentes (Sánchez, 2004), atendiendo a las actividades que se centran en la forma y a las que se centran en el contenido. Por un lado, existen actividades más enfocadas al aprendizaje expreso de aspectos léxicos o metalingüísticos (o conocimiento explícito) y, por otro, hay actividades más centradas en el contenido y en el significado, cuyo objetivo es que los alumnos practiquen con la lengua, para inferir y adquirir de forma inconsciente los contenidos (conocimiento implícito).

Dado que la creciente cantidad de materiales didácticos para la Enseñanza del Inglés como Lengua Extranjera favorece la proliferación de distintas tipologías materiales didácticos, clasificar las actividades y su potencial de aprendizaje es un asunto de especial relevancia (e.g. Criado & al., 2009, 2010, 2012; Tomlinson, 2003, 2011), pues una selección de materiales didácticos equivocada o sesgada puede acarrear consecuencias negativas sobre el producto de la enseñanza y del aprendizaje; tanto el desarrollo desequilibrado de las destrezas como el énfasis, por ejemplo, sobre la gramática en perjuicio de las habilidades comunicativas podría predefinir el tipo de conocimiento (explícito o metalingüístico, y/o implícito o más centrado en el contenido) que adquirirá el aprendiz. Todas las actividades implican estrategias diferentes para lograr objetivos concretos y, por tanto, es necesario que los profesores sepan distinguir entre tipos de actividades, según su potencial didáctico. Sería por tanto necesario establecer clasificaciones fiables basadas en una tipología de actividades extraídas de un gran banco de materiales didácticos para orientar a los profesores en la elección de estrategias didácticas enfocadas a conseguir determinados objetivos. La lingüística de corpus y sus técnicas (Sánchez & al. 1995, entre otros) puede ayudarnos a ese fin.

Para lograr el objetivo indicado al principio, se procederá de la siguiente manera: (i) escanear y digitalizar unidades didácticas completas de 16 libros de texto a fin de compilar un corpus digitalizado de actividades; (ii) aplicar programas de gestión textual como Monoconc y WordSmith para elaborar listas de frecuencia y palabras clave detectadas en las actividades (especialmente las instrucciones iniciales); (iv) analizar las listas de frecuencia y palabras clave para y determinar a qué categoría pertenecen las actividades (centradas en las formas o en mensajes comunicativos); (v) proceder luego a la categorización de las actividades docentes según su énfasis en la forma o en el contenido.

A partir de esta clasificación más genérica cabe la posibilidad de subcategorizar posteriormente otros tipos (o subtipos) de actividades más específicas.

La categorización de actividades docentes en la enseñanza de la Enseñanza del inglés como Lengua Extranjera tiene una importancia capital, ya que permitirá al profesor seleccionar con más facilidad y autónomamente aquellas actividades que mejor se ajusten a la consecución de sus fines, centrados en la forma o en el contenido, o en algunas de sus sub-variantes.

BIBLIOGRAFÍA


(49)

Gutiérrez, Marco (Universidad del País Vasco, España): A propósito del DECOTGREL: retos técnicos y metodológicos en la lexicografía técnica de las lenguas de corpus

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOPHY

La lexicografía técnica es sin duda un campo que ha sido objeto de un creciente interés en las últimas décadas y que ha invadido casi todas las áreas del conocimiento científico. Sin embargo, este desarrollo cuantitativo se ha sustentado más en el interés que las editoriales han mostrado a causa de la buena acogida que dichos instrumentos especializados han tenido en el mercado que en los propios avances metodológicos que ha experimentado esta rama de la lexicografía.

Entendemos que una excepción al estado de cosas suculentamente descrito la constituye el DECOTGREL (Diccionario Electrónico Concordado de Terminología Gramatical y Retórica Latina), elaborado por un Equipo de investigadores que lleva ya casi una década trabajando en este campo y que ha elaborado ya las dos primeras entregas de dicho trabajo (DECOTGREL (Pmin) y (Pmai).

Tanto los presupuestos teóricos utilizados para elaborar dicho instrumento lexicográfico como la disposición del material allí consignado resultan completamente novedosos, de suerte que se ha intentado optimizar el rigor científico de la recogida y disposición de los datos allí consignados y, a la vez, hacerlo compatible con unos medios que permitan una consulta rápida y eficaz de los mismos.

Entendemos que ambos objetivos resultan ser fundamentales para aquellas disciplinas que tienen como referente único (o fundamental) textos escritos. Ello hace que el método utilizado por nosotros ser aplicado en sus aspectos fundamentales a aquellos campos del conocimiento científico referenciados en lenguas de corpus.

(50)

Hasebe, Yoichiro (Doshisha University, Japan): Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks

PANEL: CORPUS DESIGN, COMPILATION AND TYPES:

This paper introduces a web-based system that was developed to conduct a full text search in a corpus of English presentation transcripts provided by TED (http://ted.com). The alpha version of the system named TED Corpus Search Engine (TCSE) is available online, currently holding data of 1,798 talks in English, which are comprised of 4,502,920 words (of 80,270 lexical items) and part-of-speech tags attached to them using Enju
English Parser. Among the core functionalities of the system are as follows:

1. The user can conduct a full text search for segments of talks that match specified patterns in TED Talks using either surface forms, lemma, part-of-speech tags, or mix of these. A “segment” refers to the basic unit of linguistic expression used by TCSE. It corresponds to a line of subtitle that is shown at a time in the presentation video.

2. The context of the segment can be easily studied both in text and video formats. On TCSE, in addition to the matched strings and lines immediately surrounding them, the whole text comprising the talk is available to use. Likewise, the video of the talk can be played either from the location where the matched segment is about to be uttered or from the very beginning.

3. With TCSE, it is easy to retrieve various types of information related either to the each segment (relative position in the talk, part-of-speech analysis of the segment, frequency of the words within, etc.) or to the talk as a whole (the title, the speaker, and the brief description of the talk).

4. In addition to the original English presentation text, TCSE is capable of dealing with translation text in non-English languages. Although Japanese is the only language available at the moment, it is potentially possible to extend the system to cover data in many other languages that are provided by TED thanks to work by volunteers.

TCSE is designed to aid linguists who need usage data of semi-formal spoken English expressions as well as language teachers who need real examples of such types of English to use in class. Even though the size of the corpus is far from comparable to some of the larger English corpora, the TED Corpus can provide examples of numerous expressions that are commonly used in the public speech situations, and the search system offers some useful functionalities, in addition to 1 - 4 above, to obtain real samples of words, phrases, and many kinds of patterns that linguists and language teachers would require in their research. For instance, the functionalities of displaying frequent n-grams (2-grams to 4-grams) and showing dispersion indices would be useful for (corpus) linguists. For language teachers, the “unique link” functionality would be helpful which presents a URL that offers direct access to a page where the text segment in question is highlighted in context and the video passage of the segment is automatically played.

In addition to introducing the design and implementation of the system, the paper also discusses its capability of dealing with languages other than English. Currently TCSE only holds translation data in Japanese, but the future extension to cover other language data could open up the possibility to use the system as multilingual parallel corpus.

(51)

Hennenmann, Anja (University of Postdam, Germany): Construction Grammar and Corpus Linguistics: The Example of con respecto a

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

The diachronic development of topic markers in Romance languages such as Spanish con respecto a ‘with respect to’ is still a research desideratum. With regard to topic markers in general only the development of French quant à has been investigated so far (cf. Combettes et al, eds., 2003). However, it has never been analysed against the background of Construction Grammar, as the framework of CxG is still not often applied in Romance linguistics (but see De Knop et al, eds., 2013).

Con respecto a is considered a construction because its form, function and even meaning “is not strictly predictable from its component parts or from other constructions
recognized to exist” (Goldberg 2006: 5). Since con respecto a is (nowadays) usually followed by a noun phrase or pronominal phrase, which as a whole can be identified as a construction, this study adopts a constructionist approach in terms of frequency and entrenchment, to the rise and development of this ‘topic marking construction’:

(a) Con respecto a la envidia que algunos de mis compatriotas me tienen [...] (Sepúlveda: Epistolario. Selección) [16th cent.]

(b) […] qué desventajas tendrá nuestra nación con respecto a la inglesa [...] (Valle Santoro: Elementos de economía política ...) [17th cent.]

(c) […] y dulzura de los versos, considerados cada uno de por sí y con respecto a la colocación de las sílabas [...] (Jovellanos: Prosa. Selección) [18th cent.]

(d) Con respecto a la comedia, sea en buen hora el espejo de la vida [...] (Larra: Artículos) [19th cent.]

(e) […] y no tienen ningún privilegio con respecto a todo el grupo. (Mex: Yucatán: 97Jun12) [20th cent.]

The work with the Corpus del Español shows that its frequency has increased over time so that the construction can be nowadays said to occur “with sufficient frequency” (Goldberg 2006: 5). Additionally, the results show that the construction con respecto de instead of con respecto a is also to be found:

(f) [...] & las cosas que contienen con respecto a algunos logos dela tierra; al cielo. (Alfonso X: Libros del saber de astronomía) [13th cent.]

(g) […] aún aumentaré que creo que él sea el que os llena, con respecto de los hombres, de debilidad [...] (Trigo: Mi media naranja) [19th cent.]

Till today both constructions con respecto a and con respecto de are co-existent. Nevertheless, the use of the latter seems to decrease – at least not significantly increase – as the percentages show in comparison to the total amount of tokens for con respecto a/de:

<table>
<thead>
<tr>
<th>Century</th>
<th>cxn – tokens</th>
<th>Of these con respect de - tokens</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>15</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>18</td>
<td>86</td>
<td>1</td>
<td>1.16</td>
</tr>
<tr>
<td>19</td>
<td>471</td>
<td>12</td>
<td>2.55</td>
</tr>
<tr>
<td>20</td>
<td>755</td>
<td>25</td>
<td>3.31</td>
</tr>
</tbody>
</table>

It is important to note that 16 out of 25 examples for con respecto de in the 20th century are retrieved from Mexican speech so that we possibly deal with a diatopically marked construction. It will also be worked with the corpus programme CREA in order to retrieve examples from the 21st century.

A further goal of the study is to describe the construction in more detail, showing its characteristics, its use and its restrictions of use, focusing in particular on its syntactic position.
Hernández-González, María Belén (Universidad de Murcia, Spain). Connotaciones socioculturales del término extranjero en los textos administrativos italianos del corpus LADEX

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

El fenómeno de la inmigración, la globalización y apertura de fronteras europeas, han determinado la erogación creciente de leyes y documentos administrativos relacionados con el ámbito de la extranjería, tanto en España como en otros países de la UE. Sin embargo, desde el punto de vista lingüístico aún no existían estudios basados en análisis de corpus sobre la evolución del lenguaje jurídico-administrativo relacionados con este campo, ni con otros cuyo cambio haya sido patente la última década. El proyecto LADEX (Lenguaje Administrativo de textos de Extranjería), desarrollado en la universidad de Murcia durante los últimos tres años y financiado por MINECO ha recopilado el primer corpus multilingüe (español, francés, inglés e italiano) integrado por documentos administrativos y jurídicos paralelos y estructurado en cinco secciones:

1) Textos normativos (divididos en: reglamentos, decretos y leyes).
2) Actos administrativos dirigidos al ciudadano (subdivididos en: informes, certificados, citaciones resoluciones).
3) Documentos informativos (folletos, informes, páginas web, memorias técnicas).
4) Actos administrativos dirigidos a la propia Administración (divididos en: circulares, declaraciones, relaciones, memorias).
5) Actos administrativos dirigidos del ciudadano a la Administración (divididos en: peticiones, solicitudes, declaraciones, comunicaciones).

La anotación del corpus LADEX en italiano ha revelado algunas connotaciones socioculturales del uso común del término extranjero, determinadas por la intención de los textos y codificadas según la colocación del término en la frase, las construcciones sintácticas que lo acompañan con más frecuencia y el valor de los modificadores. Dichas connotaciones se corresponde según lo observado con construcciones semánticas deliberadas. La presente comunicación muestra la caracterización lingüística de la persona extranjera no perteneciente a la UE desde la perspectiva de las instituciones italianas nacionales, regionales y locales entre los años 2011 y 2013, un período particularmente problemático por el incremento de los flujos migratorios en los territorios del Sur de Europa. Los resultados estudiados serán de utilidad tanto para el estudio comparado de las lenguas del corpus como para el desarrollo de instrumentos para la traducción y la mediación cultural.

Herrero Zorita, Carlos, Moreno-Sandoval, Antonio (Universidad Autónoma, Madrid) & Ueda, Hiroto (University of Tokyo, Japan): Productivity of combinations of Spanish anatomical themes with symptom suffixes based on quantitative analysis

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The following work presents an exercise regarding the level of compatibility of a series of Spanish medical themes and suffixes. The objective is to study, from a number of suffixes denoting illnesses or symptoms, which can be attached to themes or prefixes that concern
parts of the human body. For this purpose, we have extracted the frequency of each combination from a Spanish medical corpus and performed several statistical operations, which show that not all suffixes accept each prefix and vice versa. A secondary objective would be to evaluate which of these statistical operations is the most preferable regarding the presentation of the results.

Although issues regarding morphological restriction and morphological productivity have a long tradition in relation to quantitative measuring (Baayen and Lieber, 1991; Baayen, 1992, 1993; Bauer, 2001; Hay and Baayen, 2002, among others), research has been focused mainly on derivational morphology. Building on this tradition, the present work would focus on medical lexical affixation using concentration measures of frequencies extracted from a specialised corpus. We believe there are certain constraints in the usage of medical term formation regarding part of the body-symptom. Thus, the present work will analyse the productivity of a series of affixes, following the prerequisites of productivity: frequency, semantic coherence, and ability to form new words (Bauer, 2001: 20).

First of all, we have selected thirteen Spanish themes that are related to several parts of the human body, from arteri(o)- (related to an artery) to tiroid(o)- (related to the thyroid gland) and fourteen suffixes that refer to symptoms or diseases, from -algia (“pain”) to -tóxico (“poison”). Secondly, we have studied the frequency of each prefix and suffix in the corpus, and the word that is formed through each possible combination. Finally, we have performed a series of different matrix and clustering statistical tests that would allow us to observe the degree of acceptability of each affix and its productivity.

A preliminary test has shown that we can situate the productivity on a scale of different groups. For example, we can observe a group of symptoms suffixes that can only be attached to one to three very specific prefixes: -cele (“tumour” or “hernia”), -megalia (“irregular enlargement”), and -tóxico (“poison”) seem only to accept those prefixes related to the heart, the liver and the bone marrow. Also, we can find four of our symptoms suffixes attached to nearly every anatomical prefix: -itis (“inflammation”), -osis (“diseased condition”), -patía (“illness”) and -tomía (“incision”). A further analysis of their frequencies and also semantic information will allow us to draw some conclusions on their degree of productivity.

The corpus that will provide the data is the Spanish subcorpus of the MultiMedica corpus (Moreno-Sandoval & Campiños-Llanos, 2013) with more than 4 million words gathered from different medical sources. The posterior statistical procedure has been performed using the NUMEROS software developed by Ueda Hiroto.

We believe this work would prove to be useful for two reasons. First, it will provide specific evidence on the productivity of medical affixes regarding human anatomy and symptoms, useful for further work on fields such as term recognition. Secondly, it will also provide a benchmark for different statistical operations regarding matrix manipulation and clustering.


(54)

Hou, Zhide (Jinan University, China): A Critical Analysis of Media Reports on China’s Air Defense Identification Zone

PANEL: SPECIAL USES OF CORPUS LINGUISTICS

This study is based on the analysis and comparison of news articles about China’s announcement of Air Defense Identification Zone (ADIZ) through the three corpora built from English-language Chinese media, western media and Taiwanese media news reports. Applying this to the corpus linguistic technique of automated semantic tagging with a discourse-historical Critical Discourse Analysis (CDA) framework, the article demonstrated how processes such as two-word concgrams, keywords, key semantic categories and relevant concordances analysis were able to identify representations of China’s ADIZ as well as directing qualitative analysis. It further examined how ideology is reflected in this context via the media discourses in relation to China’s ADIZ.

(55)

Hoyas Solís, José Antonio (Universidad de Extremadura, Spain): Early feminist or mainstream writer? An analysis of George Moore’s use of verba dicendi in three novels.

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

George Moore has seen by some of his biographers (e.g. Hone 1936, Frazier 2000) as well as literary critics (e.g. Jaime de Pablos 2006, Smith 2006) as a committed feminist, who supported women’s right to take an active role in society outside the home, praising their human, artistic or intellectual qualities rather than their morality or the attributes more usually associated with their role as home-makers. This view has been tempered recently through an analysis of modifiers used in a corpus of Moore’s novels (Hoyas Solís 2014) in comparison with findings in corpora of other writers of the time (Bäcklund 2006). This study indicated that George Moore was a man of his time, deeply sympathetic to the cause of women. However, it did not show that George Moore’s depiction of women displays a marked bias in favour of women. This study thus made a small contribution to showing how the kind of linguistic analysis made possible by corpus linguistics can contribute to confirming or challenging the expert, though somewhat intuitive approach to literary texts offered by scholars when assessing the positioning of a writer such as Moore as regards issues such as feminism. Continuing with this research into how corpus linguistics methodology can help shed light on the attitude of a novelist such as Moore to his female characters, this presentation reports research into another area of language use in A Mummer’s Wife (1884), A Drama in Muslin (1886) and Esther Walters (1894): the choice of verba dicendi. The comparison between the verbs used to report – and evaluate – the speech of men and women in these novels contributes to Moore’s portrayal of women. The analysis of the verbs used to report the manner in which women and men take part in conversation –an activity which can be seen not just as an index of larger social activities generally but indeed the foundation of all social action– provides another way of assessing the image of women offered in these novels and on the roles they play in the social worlds they inhabit.
Bibliography

(56)
Ilamola, Cristina (Universitat de Barcelona - CUSC, Spain): ¿Será la lengua inicial, la televisión, o van a ser los amigos? Un caso de convergencia lingüística en el ámbito gramatical por contacto entre castellano y catalán.

PANEL: CORPUS AND LINGUISTIC VARIATION
No son nuevos los trabajos de base variacionista que ratifican el avance de Ir a + infinitivo (FA) frente a la forma sintética en -ré (FS), para expresar la futuridad en castellano (Troya, 1999; Almeida, 1998; Sedano, 1994, 2005). Pero, desde el ámbito del estudio de los fenómenos de contacto de lenguas, se han realizado otros trabajos que analizan la distribución de ambas formas cuando el castellano se halla en contacto con otras lenguas (Blas Arroyo, 2008; Buzón García, 2013; Ilamola, 2008, 2013), y concluyen que la FS se usa en mayor porcentaje que la FA.

Esta comunicación se inscribe en la línea de estos últimos y pretende exponer los resultados de la aplicación de nuevas variables en el estudio de este fenómeno en situaciones de contacto lingüístico. Los estudios citados, que parten de corpus orales de informantes bilingües castellano-valenciano, usan las variables sexo, edad y lengua inicial. Y concluyen que la L1 es la que justifica el empleo de una u otra forma: los valencianohablantes/bilingües iniciales prefieren la FS; los castellanohablantes iniciales se inclinan en mayor grado por la FA.

En esta ocasión nos valemos del corpus oral escolar del proyecto RESOL , constituido por entrevistas guiadas a niños de entre 11 y 13 años de diferentes municipios de Cataluña. Concretamente, se examinan las producciones de FS y FA de informantes de Mataró (Barcelona) en dos momentos distintos: en 6º de primaria y, posteriormente, en 4º de la ESO.

Dado que estos individuos pertenecen a una de las comunidades lingüísticas más bilingüizadas de Europa, esas tres variables resultan insuficientes para comprender sus prácticas lingüísticas, complejas. De ahí que ahora se incluyan la lengua familiar, la lengua con los amigos y la lengua de sus programas televisivos favoritos.
En definitiva, pretendemos demostrar que las producciones lingüísticas de los individuos bilingües están sujetas a variables sociolingüísticas mucho más complejas que la tradicional lengua inicial y que contribuyen a la comprensión de fenómenos de variación.

(57)
Ishikawa, Shin’Ichiro (Kobe University, Japan): Degree of Divergence between Spoken and Written Vocabularies as a Means for Classification of L2 Learners

PANEL: CORPUS, LANGUAGE ACQUISITION AND TEACHING

1. Introduction

Vocabularies used in speeches and writings can be different in many ways (Biber et al., 1999; Leech et al., 2001). This is also true to L2 use. However, the content of two kinds of vocabularies and the discrepancy between them are changeable according to varied conditions. In some situations, the discrepancy becomes larger, while in other situations, it becomes smaller. This suggests the possibility that the degree of divergence between spoken and written vocabularies (DDSWV hereafter) can be utilized as a means to classify varied speakers and writers, including L2 learners with different L1 backgrounds and at different L2 proficiency levels.

Although this might be promising, it is not easy to compare the two kinds of vocabularies in a reliable way. For, the content of the vocabularies can be easily influenced by varied elements including topics and genres. Unless they are controlled, it would be extremely difficult for us to say what the “difference” identified by the comparison represents.

For a reliable comparison of L2 spoken and written vocabularies, we need a database collecting L2 speeches and writings produced about the same topics and in the same conditions. Thus, the author has been in charge of compilation of The International Corpus Network of Asian Learners of English (ICNALE). This large-scaled learner corpus includes spoken and written modules and its striking feature is that the same prompts are given to speakers and writers (Ishikawa, 2013; Ishikawa, 2014). Using the ICNALE, we aim to investigate the applicability of DDSWV for learner classification.

2. Research Design

2.1 Aim and RQs

The current study aims to clarify whether DDSWV can be utilized as a means to classify varied learners of English. We also examine the spoken and written vocabularies used by English native speakers as a baseline for comparison. Thus, two research questions are posed: RQ1 Are learners at different L2 proficiency levels classified appropriately according to DDSWV? and RQ2 Are learners with different L1 backgrounds appropriately classified according to DDSWV?

2.2 Data

We use the ICNALE-Written Version 2.1 and the ICNALE-Spoken Baby Version 1.3. In order to minimize the topic influence, we use the speeches and essays about the same one topic: non-smoking at the restaurants. Concerning RQ1, we analyze the Japanese, Chinese, and Taiwanese learners at A2, B1_1, B1_2, and B2 proficiency levels (These levels are based on the Common European Framework of Reference). Concerning RQ2, we analyze the learners in four EFL countries and areas (Japan, China, Indonesia, and Taiwan) and two ESL countries (the Philippines and Singapore) as well as English native speakers.
2.3 Methodology

We define DDSWV in two ways: (a) 1- correlation values and (b) squared distances on the scatter plot (Z1 X Z2) obtained from correspondence analyses. The target is limited to the top 100 high-frequent words.

3. Results and Discussions

Concerning RQ1, it was revealed that DDSWV can function to some extent as a means for classification of learners at different L2 proficiency levels, but its reliability changes according to learner groups. Concerning RQ2, it was revealed that DDSWV does not clearly distinguish the learners in EFL region from those in ESL region.

(58)

Ishikawa, Yuka (Nagoya Institute of Technology, Japan): A Corpus-based Approach to Gender Differences in Language: Do Men and Women Write Differently?

PANEL: CORPUS AND LINGUISTIC VARIATION

Since Lakoff (1975) called attention to linguistic differences between genders, numerous empirical studies have been conducted examining linguistic features related specifically to men and women. Koppel et al. (2002) analyzed 566 texts taken from the British National Corpus to identify linguistic features more commonly used by one or the other gender. They assert that the male indicators were largely noun specifiers whereas the female indicators were mostly negation, pronouns, and certain prepositions. Newman et al. (2008) compiled a large corpus and studied gender differences in language use. They claim that women in their corpus used more words related to psychological and social processes and more verbs, whereas men discussed current concerns and used more words related to object properties and impersonal topics. However, previous studies have usually focused on texts produced in uncontrolled conditions. Therefore, we cannot deny the possibility that other factors than gender may have affected the results. Do they really interpret things differently?

This study investigates gender differences in language use in L1 argumentative essays written by men and women on designated topics under controlled conditions using a large essay corpus, the ICNALE. The results of the study presented here indicate that there are indeed gender differences in L1 language use in essay writing, regardless of their L1 is English or Japanese. Men from English speaking countries tend to use more nouns related to social economic activities to convey information or facts about the given topics, whereas women tend to use more pronouns, more intensifiers and modifiers, and words related to psychological cognitive processes so that they might convey their feelings and develop a good relationship with other people. Japanese men tend to use casual form particles and women tend to use polite form of particles at the end of the sentences. Discriminant analysis is used to classify essays into two groups according to the author’s sex and the correct classification rate is 80.0%.

(59)

Izquierdo, Marlén (University of the Basque Country, Spain): Structured big data in business: the audit report from a cross-linguistic perspective
PANEL: CORPORAL, CONTRASTIVE STUDIES AND TRANSLATION

Business practices do not escape the increasing call for transparency in today’s society. In this sense, the business community fulfills a variety of tasks, some of which take on also a linguistic dimension. Such is the case of auditing, an essential practice intended to attest the financially accurate state of a business –hence, its ethical running- through the gathering and analysis of relevant, financial, data (Flowerdew and Wan, 2010). The result of the examination carried out is officially reported in the so-called audit report (AuR). As a professional genre, it has a clear-cut communicative purpose, which is achieved through a set of rhetorical and lexico-grammatical conventions agreed upon -and expected- by the business discourse community (Bhatia, 2004).

Given that languages map out shared meanings differently, and framed in an international context of approximating cultures (Upton & Connor, 2001), this study aims to contrast the similarities and differences of the rhetorical and lexico-grammatical realization of the audit report in English and in Spanish. Data has been taken from an ad-hoc comparable corpus, ACTRES C-AuRs. Adhering to the ESP approach to genre (Bhatia, 1993; Swales, 1990), a top-down study unfolds in order to firstly account for the prototypical structure of the audit report prior to describing the typical lexico-grammatical realization of the various moves and steps of the genre in a second stage. In other words, the macrostructure of the genre is taken as the tertium comparationis on the basis of which verbalization patterns are juxtaposed and contrasted. The findings reveal that English and Spanish auditors do not seem to follow exactly the same reporting procedure –as observed in minor differences of move and step insertion within the report. In addition, while the genre is rather formulaic in both languages, differences crop up regarding the actual actions taken by the auditors, at the different stages of the auditing activity, which ultimately portrays the transparency of the auditing itself, rather than that of the company’s financial statements.

The pedagogical implications of approaching ESP from a genre view are discussed, highlighting the benefits for would-be auditors, who need be aware of how they are expected to report their activity. Furthermore, not only would the results contribute to syllabus or ESP material design, but also to accurate training of Spanish professionals who need write in English, a major means of business communication (Flowerdew and Wan, 2010).

(60)
Jorgensen, Annette Myre (Bergen University, Norway): La creación de la identidad a través del discurso narrativo en conversaciones adolescentes

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

Para definir la identidad hay que analizar los procesos de la interacción entre las personas, ya que la idea sobre uno mismo y los demás de acuerdo con Strauss (1959) se forja en esa interacción. En añadidura, la identidad de una persona marca también su modo de interactuar (Tracy 2002). En este trabajo enfocaré la interacción de los adolescentes y el modo de presentarse ellos mismos en las conversaciones informales en grupos de amigos, en eventos narrativos breves. El grupo de amigos son comunidades de habla (speech communities) en las que los hablantes adoptan los modos de hablar propios de esa comunidad (Brown/Levinson 1987). Como señala Spreckels (2009: 49) de los adolescentes: “the intensive investigation of small-scale peer group interaction certainly helps us to better understand the sometimes intricate and subtle features of youngspeak.”

Un ejemplo de la construcción de identidad de los adolescentes es el uso que hacen del
lenguaje vago. Parte de su identidad como miembros del grupo consiste en compartir el mundo que los rodean, y aluden a él con términos vagos, creándose así una identidad de “miembro del grupo”.

El análisis del lenguaje vago en los eventos narrativos breves adolescente se basa en las conversaciones del corpus COLA de la Universidad de Bergen (www.colam.org).

(61)

Jorgensen, Annette Myre (Bergen University, Norway) & Esperanza Eguia Padilla (COLA-prosjektet, Spain): Presentación de COLA, un corpus oral de lenguaje adolescente en línea

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

El Corpus Oral de Lenguaje Adolescente: COLA de la Universidad de Bergen, Noruega contiene conversaciones espontáneas, informales recogidas en Madrid (500.000 palabras), Santiago de Chile (100.000 palabras) y Buenos Aires (80.000 palabras) en línea: www.colam.org (Jørgensen & Drange 2012). COLA es un recurso que ofrece la posibilidad de investigar y transmitir en la enseñanza del español, el estilo comunicativo, la interacción, los marcadores del discurso y el léxico de los jóvenes de habla castellana a lingüístas y otros interesados en el lenguaje juvenil/oral (Jørgensen 2008).

Las conversaciones están trasladadas a formato digital y transcritas con el programa Transcriber, cuya ventaja es que sincroniza los archivos sonoros con el texto transcrrito, es decir, se puede leer la transcripción y escuchar el sonido al mismo tiempo. Las conversaciones están transcritas ortográficamente. En las transcripciones se respetan las reglas ortográficas independientemente de la pronunciación de las palabras (Drange 2009). Las conversaciones se han convertido en textos en archivos HTML enlazados a los archivos de sonido WAV, trasladados al formato de Corpus Workbench a Internet (Hofland, Jørgensen et al. 2005) permitiendo búsquedas por sexo, edad, clase social.

Queremos exponer el proceso de construcción del corpus COLA, con resultados de la investigación y la rentabilidad de este corpus discursivo en la didáctica del español en el aula.

(62)

Kano, Makimi (Kyoto Sangyo University, Japan): Revealing Factors Affecting Learners’ Sense of “Difficulty” in Extensive Reading through Reader Corpora

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

Extensive Reading (ER) is widely accepted as an effective assignment in many university English curricula. Graded Readers (GR) specifically designed for learners of English as a foreign language are generally used in ER programs. Their vocabulary is controlled, and their expressions and contents are simplified so that elementary and intermediate students of English can read extensively. However, the authenticity of GR has long been debated (cf. Claridge 2005, Swaffar, J. 1985, Honeyfield, J. 1977). Clafin (2012) claims that “youth literature should be part of the choice at all levels” so that it can be “a bridge to native speaker literature.” Recently, Youth Readers (YR), written for native speaker children, are also used in ER programs. However, students often find YR more difficult even when categorized as being at the same level as GR. Though YR have been put into difficulty levels and used in many ER programs, the differences between GR and YR had
never been objectively analyzed.

To find out what makes learners of English feel that YR are difficult, two kinds of corpora, ER and YR, have been compiled. Each of these corpora contains about 300,000 words from readers chosen as most popular among 1st year students taking compulsory English courses at a university in Kyoto. The GR and YR reading levels ranged from 1 to 5 (≈CEFR A1-B1), and the word counts in each level are well-balanced so that two corpora can be compared by level. The data from the two corpora were analyzed using various corpus tools and vocabulary lists: the functions of Word List, Keyword List, Clusters/N-grams of AntConc, AntWordProfiler with BNC/COCA word family lists and General Service List, Flesch Reading Ease and Flesch-Kincaid Grade Level etc.

Comparing the YR corpus with the GR corpus, no significant differences were observed in readability scores, and their average word lengths and words per sentence were similar, but some interesting differences were observed. The YR corpus contains twice as many word types and a lower percentage of the basic 1000-word-level vocabulary, and shows a rapid increase of vocabulary level as the reader levels escalate. In YR, there is also a higher percentage of passive sentences and complicated verb phrase structures, such as auxiliary verb + perfect (e.g. must have been) and perfect + progressive (have been running). Some basic words, such as even, if, been and around, which are often used as one or two fixed or limited phrase in GR, have much higher frequencies and more usage varieties in YR. Some colloquial expressions or slang such as icky, twerpy, jiffy are also exclusively used in YR. At the same time, more descriptive sentences are found in YR while GR contains a lot of you and I centered conversational sentences. These characteristics of YR may be considered as factors that affect learners’ comprehension.

Reference
Claflin, M. (2012). Bridging the Gap Between Readers and Native Speaker Literature. Extensive Reading World Congress Proceedings, 1, 156-159, retrieved June 6, 2014,


(63)
Khakimov, Bulat, Galieva, Alfiya M. (Kazan Federal University, Russian Federation), Suleymanov, Dzhavdet (Tatarstan Academy of Sciences, Russian Federation) & Nevzorova, Olga (Kazan Federal University, Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Russian Federation): Corpus-based Turkic lexicography: applications of Tatar National Corpus

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The development of Turkic linguistic studies during the past years has been marked by deepening of the theoretical basis of linguistic research, the increasing emphasis on new directions and challenges of modern linguistics, including applied linguistics. In general, the corpus technology is getting more and more wide application in various aspects of modern linguistics. As a result, corpus-based dictionary compilation became a common method in lexicography. However, corpus-based lexicology and lexicography of the Tatar language, which belongs to the Turkic group and is spoken mostly in Russian Federation by over 5 million people, still remains virtually undeveloped area.

Despite the fact that Tatar lexicography has a long history and traditions, until today it could not take into account the actual distribution of linguistic units and all their senses in
speech. Selection of units for dictionaries has been carried out on the basis of dictionary compilers’ linguistic intuition, which is subjective in many respects. Even such a relatively modern dictionary as The Tatar explanatory dictionary (Kazan, 2005) has been compiled on the basis of hand-picked card index examples. Insufficient attention have been paid to syntactic factors influencing the meaning (word combinations, verb diathesis, order of words, etc).

This research focuses on the problems of development of the Turkic and Tatar corpus-based lexicography. What are the capabilities of development? How can we harmoniously combine traditional achievements and modern technologies? What future Tatar dictionaries will look like?

In our study we determine and evaluate possible ways to improve the structure and content of the Tatar dictionaries using the corpus data. Main tasks in this aspect are searching for words have not been included in the dictionaries; searching for the typical collocations and quotations to the already known senses of the words, as well as for new collocations; searching quotations to senses that are not illustrated by quotations in the dictionary; searching for the senses of words which have not been included in dictionaries; searching for new senses of words; improvement of the current definitions of words; analyzing the meanings of synonyms, etc.

In a corpus-based lexicographic study it is important to adequately recognize the relevant information in the context. Such information allows to choose the dominant sense of the word, as well as to solve the problem of distinction between polysemy and homonymy. Various criteria of differentiation of polysemy and homonymy are widely known and discussed in the literature: etymological analysis, difference in inflectional paradigms, various derivatives, synonym substitution method, translating words into other languages, etc. From the corpus point of view analysis of the context and lexical combinability looks the most appropriate way.

As lexicographically relevant important features of the word context in corpus, we analyze frequency and word combinability. On the example of group of synonyms and another lexical group of basic color names we compared the current non-corpus dictionary entries with the empirical data from the Tatar National Corpus in order to evaluate distinction of the word senses and the accuracy of definitions. By the method of calculating the frequency of headwords and collocations, and taking into account the taxonomic classes and some other features of the collocates, we propose updated dictionary entries and detect new actual collocations. Requirements to the corpus annotation system from the lexicological and lexicographical point of view are also discussed.

References
Kilgarriff, A. How dominant is the commonest sense of a word? Paper presented at the Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 3206, 2004, P.103-111.
Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013) / Ed. Chelo Vargas-Sierra, P.68-74.


The explanatory dictionary of Tatar language, Kazan, 2005 (In Tatar).

(64)

Khakimov, Bulat, Salimov, Farid I. (Kazan Federal University, Russian Federation) & Ramazanova, Dariya B. (Tatarstan Academy of Sciences, Russian Federation): Building dialectological corpora for Turkic languages: Mishar dialect of Tatar

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

Electronic corpus is an effective method to store, preserve and investigate dialects. Corpus-based dialectological studies represent the relatively new field in modern Turkic and Tatar linguistics. Along with the problem of “linguistic ecology” and preservation of linguistic variety, it is valuable for historical linguistics, typology and more. While written literary corpora of Turkic languages develop actively only the first steps are made in building corpora of dialects.

The purpose of this study is to discuss problems of dialectological corpus annotation, dialectological database compilation, and integrated resources for corpus-oriented and computational dialectology. The research is based on the Electronic Atlas of Tatar Dialects (http://atlas.antat.ru), which is the geoinformational web resource about dialects of Tatar language spoken mostly in Russian Federation by over 5 million people. The Atlas consists of more than 200 maps and describes territorial distribution of different phonetic, lexical, morphological and syntactical phenomena.

In 2012 we started to develop its textual extension – the corpus of Mishar dialect. Mishar is one of the main dialects and it is actively used in oral communication. The database of the corpus includes texts recorded since 1950 until nowadays and consists of about 50000 words. Dialect texts are morphologically annotated and glossed, and they are classified according to the special set of metatags. The part of the texts is accompanied by English translation. The necessary meta-information in the corpus is represented by a detailed dialectological tagset, which contains information about the dialect, the place and time of recording, the informant and subject/genre characteristics of the text.

Dialect texts in the Mishar corpus come from different sources. Some of them were collected during dialectological expeditions of the Tatarstan Academy of Sciences; another part comes from the collection of Moscow State University. And finally, it includes earlier recordings of the Soviet time, which were published in several compilations. These earlier published texts were scanned and recognized using OCR technology with adaptation to dialectological transcription character set.

From the point of view of implementation, our Mishar corpus is an indexed set of word forms. Indexes determine to which sentence and text a particular token belongs, and each token has its grammatical annotation. Grammatical annotation in our dialect corpus is based on the model of the Tatar literary language and it is consistent with commonly used typological terminology and glossing rules. In order to annotate specific dialectal grammatical phenomena, additional tags were developed. The user can specify a set of grammatical features, using a special interface.

The corpus also includes a variety of integrated resources, for example, dictionaries containing information about the tags appeared in annotation in each subdialect, providing a comparative view on peculiarities of grammatical inflection in different
subdialects within the Mishar dialect. Another special resource is the corpus-based dictionary of dialectisms. It contains information about the texts and sentences in which the dialectism appears. The dictionary also includes the literary equivalents of the dialectisms, their phonetical variants and more. This dictionary is associated with the corpus, so one can select a word in the dictionary and easily find examples from the corpus.

As a further development of the corpus of Mishar dialect we plan to increase the amount of text base, provide more detailed annotation, and implement additional integrated resources.

(65)

Kruse, Mari (University of Tartu, Estonia): En busca de influencias interlingüísticas: el corpus de trabajos finales de los estudiantes de filología hispánica en la Universidad de Tartu.

PANEL: CORPUS, LANGUAGE ACQUISITION AND TEACHING

En el campo de enseñanza y aprendizaje de lenguas, la cuestión de si las lenguas pertenecientes al repertorio de una persona plurilingüe (alguien que es capaz de comunicarse en tres o más lenguas) se interactionan entre sí había suscitado mucho debate. Últimamente, no obstante, el interés se dirige más hacia investigar cómo se produce esta interacción y cuáles son sus resultados, ya que su existencia es evidente. Según modelos recientes como el Modelo Dinámico de Plurilingüismo de Herdina y Jessner (2002) se propone que la adición de cada nueva lengua aumenta la cantidad de variables involucrados en el sistema complejo plurilingüe y produce nuevas variaciones.

Uno de los mecanismos más fundamentales para la interacción interlingüística es la transferencia, término aquí destinado a un uso amplio que describe el intento de servirse de rasgos lingüísticos y no lingüísticos conocidos para obrar en situaciones desconocidas. Ringbom (1986), pionero en estudios plurilingües, indica que la distancia percibida entre lenguas es un factor clave en transferir rasgos gramaticales y léxicos. Si percibimos la lengua que estamos aprendiendo o utilizando como distante y diferente en comparación con nuestra primera lengua, se disminuye la disposición para transferir lo que le es propio, mientras que otra lengua conocida puede percibirse como más oportuno para buscar equivalencias.

En este contexto se encaja el corpus recopilado en el Departamento de Filología Hispánica de la Universidad de Tartu. Se trata de un corpus de español estudiantil escrito, consta de los trabajos finales de grado y máster. Cuenta con 114 redacciones de diversa extensión, del registro académico, por lo cual se les puede exigir un uso ponderado y normativo de la lengua. Sus autores asimismo han respondido a una encuesta sobre su trasfondo lingüístico. Para la gran mayoría de los encuestados (94,23%), la lengua materna es el estonio, una lengua ugrofínnea que apenas se habla fuera de nuestro país. Por otra parte, en el momento de empezar los estudios del español ya conocían otras lenguas indoeuropeas. Ante otros se destaca el inglés, lengua en que contaban con una alta competencia; el 70% de los autores encuestados incluso lo consideraron la lengua extranjera dominante en el repertorio personal. Dadas las diferencias intrínsecas entre el estonio, el español y el inglés, la hipótesis de distancia percibida de Ringbom permite suponer que en su uso del español se pueden trazar influencias del inglés más bien que las del estonio. Ringbom asimismo postuló que las influencias de la lengua materna distante se limitan al nivel sintáctico y morfológico, mientras para prestar unidades léxicas se puede servir de cualquier lengua conocida, sobre todo de la que se percibe como más cercana. En suma, se pretende ver en qué extensión se pueden detectar en el corpus
A preliminary analysis of a small number of texts has provided a tentative list of rhetorical

Online cheese descriptions follow specific textual conventions which make them recognizable as belonging to a particular subgenre. These conventions imply a common overall structure where all the texts contain a similar arrangement of purposeful communicative units determined by the context of use. Several authors have proposed ways of describing the different functional units within texts that identify them as belonging to a particular genre or subgenre, including the typical linguistic features associated to each unit (Bhatia 1993, 2004; Swales 1990, 2004; Biber et al. 2007). Swales’ move-step method has been used to establish the rhetorical structure of online cheese description in this study.

A corpus-based methodology has been used here for extracting the relevant information to produce a writing tool. The corpus consists of 150 cheese descriptions in English, in all cases dealing with a wide range of cheeses produced in the UK. All the texts were downloaded from websites of either cheese manufacturing companies or more general websites describing different types of cheese. The corpus contains all in all 23,089 words, with an average number of approximately 154 words per text.

A preliminary analysis of a small number of texts has provided a tentative list of rhetorical

Bibliografía:


Labrador, Belen & Ramón, Noelia (University of León, Spain): ‘Perfectly smooth, creamy and full flavoured’: Online cheese descriptions

PANEL: SPECIAL USES OF CORPUS LINGUISTICS
This paper presents a macro and micro-linguistic corpus-based analysis of online cheese descriptions in English. Nowadays, in the region of Castille and León, a number of small companies devoted to tourism and the manufacturing of food products are interested in internationalizing their services and, thus, expanding their trades to other countries. This implies a growing need for linguistic services, not only direct translation and/or interpreting services, but also services involving assistance in professional writing for various purposes. The ACTRES project currently in progress at the University of León, Spain (http://actres.unileon.es), aims at meeting this need by building software for professional writing in a number of different fields, including wine tasting notes (López-Arroyo and Roberts, forthcoming), heritage recipes, herbal teas, rural accommodation, online advertisements (Labrador et al. 2014), and others. In the same line, the present study aims at providing a detailed account of online cheese descriptions, to help Spanish-speaking professionals in the dairy industry to write this specific text type.

Influencias de dos lenguas, del inglés y del estonio, y cómo se las puede clasificar y analizar. Como el corpus cuenta con textos redactados a lo largo de 18 años, todo el ciclo de vida del departamento de hispanismo en la Universidad de Tartu, y por autores de diferentes niveles de competencia lingüística (grado y máster), también es posible considerar el variable de competencia y si en el nivel diacrónico se pueden percibir cambios cualitativos en la redacción de los trabajos finales.
tags to be used in the process. These labels will be employed to tag the texts with an ad-hoc tagger, which will later enable us to extract concordances in particular moves, steps or sub-steps. By observing the concordance lines, the specific phraseology typical of a particular move or step is thus easily retrieved. A total of 9 different moves, some of them with steps were identified in cheese descriptions, including tags such as the geographical and historical provenance of the cheese, the type of milk and rennet used and serving suggestions like food and wine pairing. A detailed analysis will be carried out to obtain the most relevant lexico-grammatical elements contained in each move and step to produce a number of 'model lines' which may function as suggested phrases for the writing of online cheese descriptions. A further step will be the inclusion of a semi-specialized glossary.

The final product obtained from the research will be a computer tool designed to assist in the writing of online cheese descriptions in English, providing the relevant rhetorical and lexico-grammatical information for this particular text type.

References


(67)

Laippala, Veronika, Kanerva, Jenna, Missilä, Anna & Ginter, Filip (University of Turku, Finland): Syntactic ngrams as keystructures reflecting typical syntactic patterns of corpora in Finnish

PANEL: CORPUS AND LINGUISTIC VARIATION

Syntactic ngrams are little subtrees of dependency syntax trees, i.e. combinations of words related in a dependency syntax analysis. While a traditional ngram is composed of words following each other, a syntactic one consists of syntactically related words and their functions. Words in a syntactic ngram do not necessarily follow each other linearly.

This presentation uses syntactic ngrams to study typical syntactic structures of texts and syntactic differences between them. The aim is to apply the concept of keyness (Scott & Tribble 2006) to unlexicalised syntactic ngrams, i.e. syntactic ngrams with the actual words removed. Keyness refers to statistically meaningful differences between texts (Scott & Tribble 2006) and is most often studied via keywords, i.e. words that are statistically more or less frequent than would be expected (Scott & Tribble 2006). However,
unlexicalised syntactic ngrams offer the possibility to concentrate on syntactic patterns typical of certain texts. By removing the actual words, the syntactic ngrams extend the level of description beyond individual words to sequences of syntactic elements (see Ivaska 2014 for a similar approach with morphological forms).

The syntactic information attached to the words of syntactic ngrams deepen the information provided about the context of a given word. Therefore, they have been applied in computational linguistics e.g. to methods that would traditionally concentrate on the linear context of a word (see Sidorov et al. 2013). Collections of syntactic ngrams have been published for English (Goldberg & Orwant 2013) and for Finnish (Kanerva et al. 2014), and the code to produce syntactic ngrams for Finnish is also publically available1.

1 https://github.com/jmnybl/syntactic-ngram-builder
2 https://github.com/TurkuNLP/Finnish-dep-parser

This ongoing work applies the syntactic ngram generation pipeline for Finnish to four corpora representing various registers and topics: Internet discussion forum texts concerning the attitude of Finnish social workers towards their customers (50,998 words), Internet discussions following the news articles and editorials in a major Finnish newspaper (102,952 words), news articles from a newspaper and an online magazine (25,353 words), articles from the Finnish Wikipedia (44,510 words) and Finnish literature (98,525 words). The corpora have automatical syntactic analyses provided by the Finnish Dep Parser2 following the relatively detailed Stanford Dependencies dependency scheme (de Marneffe & Manning 2008) with 48 dependency types. This allows a very detailed description of different syntactic patterns.

The key structures chosen into the analyses are biarcs composed of three syntactic elements and triarcs composed of four elements, quadarcs composed of five elements being too scattered and arcs composed of two elements too general. The lexical information and the morphological features, such as case and number, are deleted, but the part-of-speech categories are kept.

The first results show that syntactic ngrams reflect both the contents and the syntactic characteristics of the corpora and thus offer useful information and pointers to more detailed investigations, obviously needed in order to understand the phenomena behind the findings. Similarly to lexical n-grams over individual words, quadarcs are more informative than triarcs, although the tendencies are similar: adverbs as well as comparative and copular constructions are overrepresented in both of the Internet discussion corpora whereas the literature corpus is characterized by sentence complexity and coordinations. Names and compound nouns often used for instance with titles and explanatory appositions seem to be typical of news and Wikipedia.

References:


Kanerva, J.; Luotolahti, J.; Laippala, V.; Ginter, F. 2014. Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish. Proceedings of the Sixth International Conference
The data needed for the study was extracted from the Wells Corpus. Wells

In order to perform the research and the analysis the browsable database of CHILDES was consulted. The data needed for the study was extracted from the Wells Corpus. Wells

Baltic HLT. 2014.


(68)

Lakomski, Tomasz (NKJO Chojnice, Poland): The Acquisition of Interjection oh in Early Childhood Observed on the Basis of CHILDES Database

PANEL: CORPORAS, LANGUAGE ACQUISITION AND TEACHING

The research undertaken aims to shed more light on the acquisition of interjection oh. Though they pervade everything we say, interjections have been neglected in linguistic research due to their ambivalent nature as well as their ambiguity and difficulty to classify them according to the tools of traditional grammar. The problem has been recognized by James (1973) and then Stange (2009) who have written that interjections rarely appear in linguistic research, and if they do appear, they are treated as something less than deserving scientific interest:

Interjections are among the most little studied of language phenomena; as one looks for references to them in the works of linguists, one is struck by the fact that are very rarely mentioned, and where they are mentioned, it is usually only briefly and cursorily (James 1973: 1).

The lack of interest in interjections seems unjustified in the light of plethora of functions they are used in. Before children master ways of communicating with their environment using precise words, they express themselves with short, easy to produce sounds. These sounds, more often than not, take the form of interjections which help children fulfill their basic needs. Interjections are used by children to express their direct overt emotions and to operate on the environment. Sometimes children use an interjection as a reaction to pain, and at other times it is used to manipulate parents to get what they want. Interjections as one of the earliest phenomena present in children’s acts of communication give a perfect insight into child early language development. Which is not to say, that adults do not use them. They use them, though to different degree and in different contexts

In the study undertaken, the interjection oh mentioned earlier, will be analyzed with reference to many pragmatic functions which it may perform. This is one of the most frequent interjections in the children’s linguistic repertoire, and the one that children acquire relatively early. It will be observed when it is performed as a reaction to the novelty exerted on the child by the environment in which it grows. Specifically it will be taken into consideration and analyzed when it is used to suggest lack of agreement, excitement, lack of interest, disappointment, attention gainer and a reaction to something unpleasant. Moreover, oh will be scrutinized when it is used by a child with combinations with dear, God, well, no and but.

In order to perform the research and the analysis the browsable database of CHILDES was consulted. The data needed for the study was extracted from the Wells Corpus. Wells
Corpus offers day-to-day record of interactions between children and their environment. Search for desired strings, containing the interjection oh will be conducted with the use of statistical programme CLAN.

On a more general plane, the approach to interjections advocated in the research might be characterized as usage-based. Usage-based Theory is the concept borrowed from cognitive linguistics. It recognizes individualized approach to authentic linguistic behaviour of children as a part of their general cognitive development. Theoretically present research is concerned with first language acquisition in cognitive and pragmatic vein with polysemic approach to interjections used as upheld by Anna Bączkowska (Bączkowska 2011).

References:
Stange, U. 2009. The Acquisition of Interjections in Early Childhood. Diplomica Verlag GmbH

(69)
Laso, Natalia Judith (University of Barcelona, Spain) & Suganthi, John (University of Birmingham, United Kingdom): The use of a lexical database (SciE-Lex) to assist the production of biomedical discourse

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Research has demonstrated that it is challenging for non-native speaker (NNS) writers to acquire phraseological competence in academic English and develop a good working knowledge of domain-specific collocational patterns. This project aims to explore if SciE-Lex, a powerful lexical database of bio-medical research articles, can be exploited by NNS writers to enhance their knowledge of collocations in bio-medical English writing.

This contribution will present the challenges associated with collocation for NNS writers, reflect on the benefits of a lexical database and evaluate a pedagogic approach to helping NNS writers. It will specifically report on a writing workshop conducted for these medical researchers in April 2014. The workshop involved medical researchers working on drafts of their writing using SciE-Lex. The workshop also provided an opportunity for local scholars to form a network to support their publication process.

While there is a move to challenge the international academic community to support local publications in languages other than English, this contribution reports on how we hope to fill an immediate need for our Spanish medical researchers who have to publish in English in international journals.


Bibliography


(70)

Leroyer, Patrick (University of Aarhus, Denmark) Exploiting an oral corpus for lexicographical purposes: the case of OENOLEX

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The Oenolex dictionary project (Leroyer 2011, 2013a and 2013b, 2014; Leroyer and Høy 2014; Leroyer and Gautier 2014) is addressed to the compilation of a pragmatic, online dictionary of wine tasting and is managed as an international lexicographic co-operation between the University of Burgundy in Dijon (France) and Aarhus University (Denmark). The project is commissioned by the BIVB, the branch organisation of the Burgundy wine industry in France, and the French Region of Burgundy. The goal of the BIVB is to implement an information tool aimed at the promotion of Burgundy wines. Lexicographically speaking, the decision was made to develop a lexicographic information tool aimed at the communicatively- and cognitively-oriented information needs of two categories of core users: wine experts responsible for the development of wine tasting courses at the Burgundy wine school and for the marketing communication of Burgundy wines on the one hand, and students taking wine tasting courses at the Burgundy wine school on the other hand.

The Oenolex dictionary project was initiated back in autumn 2013, and so far the conceptual phase and the data acquisition phase have been completed, while data processing is still in progress. The experts of the BIVB have been closely involved in all phases, as they have decided on the specifications of the dictionary concept, and participated actively to the generation of lexicographic data from a corpus of BIVB internal documents and from an oral corpus of recordings of authentic wine tasting interactions between teachers (wine experts) at the Burgundy wine school, and students (non- or semi-experts) taking these courses.

In this article, I will outline the methodological issues related to the design and exploitation of an oral corpus for lexicographical purposes. So far, lexicography has mainly taken advantage of written corpora, and oral corpora are seldom exploited. I will also argue, on the basis of the Klosa model of lexicographical phases (2013) that experts from the field of specialised communication and knowledge covered by the dictionary are truly needed in all phases of the dictionary process. I will finally argue, in line with Fuertes
Olivera 2012 and 2013, that specialised dictionaries can advantageously make use of mixed methods when compiling and exploiting the empirical base of the dictionary. The lexicographic team should cooperate with a panel of experts, at least to check on and validate definitions of specialised items from the lemma list that have been selected and crafted by the lexicographer. In today’s world, words can no longer be separated from things in the world, and expert knowledge is needed in almost all phases of the lexicographic work flows. In other words, specialised or not, lexicography is a truly cooperative and interdisciplinary discipline taking advantage of mixed methods of its own to adapt, in every single case, the treatment and the presentation of the linguistic material to the genuine information needs of its intended users in the foreseen use situations.

Literature


PARECE QUE ESTA "SOBRESUFICIACIÓN" SE VINCULA A UNA PERCEPCIÓN POR PARTE DE LOS HABLANTES TÉRMINOS SILÁBICAMENTE EXTSESQUIPEDALISMO, UN GUSTO POR LA CONCATENACIÓN DE SUFÍJOS Y LA CONSiguiente CREACIÓN DE TÉRMINOS SLÁBICAMENTE EXTENsoS Y, POR OTRA PARTE, INNECesarios EN LA MAYORÍA DE LOS CASOS. PARECE QUE ESTA "SOBRESUFIJACIÓN" SE VINCULA A UNA PERCEPCIÓN POR PARTE DE LOS HABLANTES DE QUE ESTOS SON, FREnte A SUS EQUIVALENTES MÁS SENCILLOS, TÉRMINOS ESPECIALIZADOS.

(71)
Lindemant, David & San Vicente, Iñaki (University of the Basque Country and Elhuyar Foundation, Spain): Building corpus-nbased frequency lemma list

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

This paper presents a simple methodology to create corpus-based frequency lemma lists, applied to the case of the Basque language. Since the first work on the matter in 1982, the amount of text written in Basque and language resources related to this language has grown exponentially. Based on state-of-the-art Basque corpora and current NLP technology, we develop a frequency lemma list for standard Basque. Our aim is two-fold: On the one hand, to propose a primary Basque lemma list for a bilingual dictionary that is currently being worked on at UPV-EHU, and on the other, to contrast existing Basque dictionary lemma lists with frequency data, in order to evaluate the adequacy of our proposal and to compare lemma lists with each other.

(72)
Llanos Casado, Laura & Villayandre Llamazares, Milka (Universidad de León, Spain): Análisis de los procesos de creación de neologismos en la prensa de España y América recogida en el CORPES XXI

PANEL: CORPUS AND LINGUISTIC VARIATION

El estudio de la variación lingüística no puede dejar al margen la reflexión en torno a una de las áreas en las que más se evidencian las diferencias entre el español peninsular y el español de América. Nos referimos a la morfología léxica y, en concreto, a la creación de neologismos mediante procesos de composición y derivación.

Cualquier lectura de textos americanos despierta en el lector español un prurito contrastivo en lo concerniente a la morfología derivativa que indica de inmediato que la divergencia es mayor que la descrita en los manuales y gramáticas sobre el español (López García 2000: 28). Abordar estudios de este calibre es más sencillo desde el momento en que contamos con corpus lingüísticos que nos permiten acotar la búsqueda atendiendo a criterios lingüísticos y extralingüísticos. En concreto, para el estudio de la neología, nos serviremos del Corpus del español de siglo XXI (CORPES XXI) que recoge más de 180 millones de formas extraídas de textos precedentes del período comprendido entre 2001 y 2012.

Acotaremos la búsqueda centrándonos en textos de prensa, pues consideramos que son los medios de comunicación los grandes difusores (y, en ocasiones, creadores) de términos. Realizan una labor de normalización y homogeneización lingüística más profunda que instituciones de prestigio como la RAE y su impacto social condiciona, en gran medida, las normas lingüísticas del español. Además, seleccionaremos textos entre los años 2008 y 2012, esto es, los cuatro últimos recogidos en CORPES XXI; pues la neología, fenómeno que se estudia a corto plazo, así lo aconseja.

Por otra parte, la estrecha relación entre el lenguaje político y la prensa se manifiesta en la mayor frecuencia de uso de términos archisilabos. Se trata del denominado sesquipedismo, un gusto por la concatenación de sufijos y la consiguiente creación de términos silábicamente extensos y, por otra parte, innecesarios en la mayoría de los casos. Parece que esta "soresufijación" se vincula a una percepción por parte de los hablantes de que estos son, frente a sus equivalentes más sencillos, términos especializados.
Trataremos, igualmente, de demostrar que, tal y como señalará García-Medall (1997, 111), la mayor parte de los procesos morfológicos en que intervienen lexemas patrimoniales conforman reglas de formación de palabras (RFP) adscritas a la gramática del español. Así, la formación de los neologismos documentados es paralela a la de otras palabras ya atestiguadas y, por tanto, las RFP utilizadas son productivas en la actualidad, tal y como lo han sido anteriormente.

En definitiva, con este trabajo pretendemos poner de relieve cuáles son las tendencias en la formación de neologismos, sirviéndonos para ello de los comodines y criterios de búsqueda que ponen a nuestra disposición los corpus del español. En concreto, y por ser el corpus que ha documentado las palabras del nuevo Diccionario de la lengua española, nos limitaremos al CORPES XXI. La aproximación a este interesante y amplio tema se hará mediante la selección de los ejemplos más ilustrativos.

(73)
López Arroyo, Belén (University of Valladolid, Spain) & Roberts, Roda P. (University of Ottawa, Canada): How specific wine tasting descriptors are?

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY
Wine tasting notes constitute a specialized genre in the field of Oenology, with their own rhetoric and language. However, the language of wine tasting notes is by no means as specialized as that of most other specialized genres. Indeed, while there are dozens of terms used to describe and evaluate wines, there are only a limited number of words that are used exclusively or primarily for describing taste.

In this study, we will first examine how wine literature attempts to analyze the descriptors used in wine tasting notes. We will then, based on a comparable corpus of 700 tasting notes per language, study a number of common wine descriptors in English and Spanish in the context of the nouns that they collocate with, in order to determine how specific or general these descriptors are in their use and meaning. On the basis of their collocability, we will categorize the descriptors into three categories and then analyze the meaning components of those descriptors that fall into the most general category.

(74)
López Mateo, Coral & Olmo Cazevieille, Françoise (Universidad Politécnica de Valencia, Spain): Recopilación de textos para la elaboración de un corpus especializado en el ámbito de la bioquímica: aspectos teóricos y metodológicos

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY
En la actualidad, es prácticamente impensable realizar un estudio lingüístico sin recurrir a un corpus. En función del tipo de investigación que queramos emprender, recopilaremos una serie de textos basándonos en unos criterios preestablecidos (tipo de documento seleccionado (de divulgación, de investigación, apuntes, programa de asignaturas, et.); autoría responsable de los contenidos, etc.) que conduzcan a un estudio lingüístico fiable y de calidad. En nuestro trabajo expondremos el proceso de recopilación de textos especializados en el ámbito de la bioquímica en lengua alemana y la elaboración del corpus. El hecho de que, por un lado, la mayoría de textos en bioquímica se publiquen en lengua inglesa y que, por otro lado, se trate de un ámbito multidisciplinar, dificulta la compilación de textos y, en consecuencia, el propio diseño del corpus. Basándonos en la
estructura conceptual del dominio estudiado, definiremos nuestro proyecto y aportaremos unos criterios que garanticen un corpus textual representativo del subcampo seleccionado y que faciliten después la extracción de los términos especializados (Cabré, 1999; Adelstein, 2004).

Referencias:

(75)
López Santiago, Mercedes (Universidad Politécnica de Valencia, Spain): El Diccionario Multilingüe de Turismo: génesis de un diccionario en línea basado en el corpus COMETVAL

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY
A pesar de la actual crisis económica, el sector del turismo continúa siendo uno de los ámbitos más activos en numerosos países. El incremento de intercambios internacionales en el sector turístico favorece los contactos entre profesionales y neófitos; entre lenguas y culturas diferentes. En este contexto, la elaboración de un diccionario multilingüe de turismo representa, por una parte, una oportunidad para contribuir a un mayor conocimiento de este sector; y por otra parte, un reto ante la complejidad de esta empresa. Para realizar este proyecto, hemos compilado un corpus de turismo, al que hemos llamado COMETVAL (Corpus multilingüe de Turismo de la Universitat de València) porque creemos como Alonso Ramos (2009:1191) que “es impensable actualmente abordar una empresa lexicográfica sin apoyarse en un corpus”. Como Martín Herrero (2009:1031) y Marcos Marín pensamos que la Lingüística de Corpus puede aplicarse a diversas parcelas lingüísticas, entre ellas la elaboración de diccionarios. Según Teubert (2009:186), el corpus es una “collection, réglée par des principes, de données du langage empirique, de textes (ou de fragments de textes), qui sont des échantillons d’un discours donné, dotés en conséquence d’une valeur représentative”. En esta línea, consideramos, como Rastier (2002:2), que un corpus debe “être aimé”, es decir, “s’il ne correspond pas à un besoin voire un désir intellectuel ou scientifique, il se périmé et devient obsolète ». En nuestro caso, el corpus COMETVAL se identifica con todas estas aseveraciones. Durante casi tres años hemos ido compilando este corpus con numerosos documentos, en español, inglés y francés, procedentes, en casi su totalidad, de páginas web de hoteles ubicados en los siguientes países: España, Francia, Canadá, Reino Unido, Estados Unidos y de América Latina (Argentina, Chile, Costa Rica, México, Perú, República Dominicana y Venezuela). Además de estos documentos, el corpus COMETVAL contiene otros escritos sobre turismo, tales como blogs y redes sociales, revistas, normativas y legislación, así como documentación e información de web de agencias de turismo. En total, más de 8 millones de palabras, alrededor de dos millones y medio por lengua, conforman este corpus. En esta comunicación, explicaremos la génesis del Diccionario Multilingüe de Turismo (español-francés-inglés). Esta obra muestra la particularidad de estar compuesta por tres diccionarios monolingües interconectados por medio de hipervínculos. En cada uno de estos diccionarios, tras la definición de cada entrada, se incluyen ejemplos, colocaciones, sinónimos, palabras relacionadas y equivalentes en las otras lenguas, procedentes del corpus COMETVAL. Tras esta presentación, nos centraremos en las unidades léxicas, en
francés, seleccionadas para su inclusión en el diccionario, con el fin de llevar a cabo un análisis semántico y morfológico de las mismas. Gracias a este análisis, comprobaremos además las divergencias y similitudes entre el léxico empleado en los textos franceses y canadienses en lengua francesa, describiendo tanto las unidades simples como las compuestas. Consideramos que el Diccionario Multilingüe de Turismo, que será de acceso libre en Internet, constituye una herramienta de consulta apropiada para conocer y estudiar el léxico especializado de los hoteles, en varias lenguas (español, francés e inglés), tanto para estudiantes, profesores y traductores como para los profesionales del turismo.

(76)

Losey León, María Araceli (Universidad de Cadiz, Spain): Corpus-based contrastive analysis of keywords and collocations across sister specialized subcorpora in the Maritime Transport field

PANEL: SPECIAL USES OF CORPUS LINGUISTICS

This corpus-based study aims to identify the keywords and collocation strengths (Gries, 2013) in different text types across sister specialized dedicated subcorpora in the maritime transport field. Within this ESP area, which has been given little attention to date from a corpus linguistics approach, the contrastive analysis will be primarily intended to determine the frequency and coverage of register or style (Leech, 2001), that is, the extent to which a terminological unit is likely to occur in different text types of a specialized corpus and whether it is possible to extract the specific distinct terms in each of the subcorpora under study despite their close ties in shared semantic fields. The designed workplan for this empirical study relies on putting differences into perspective alongside similarities (Baker, 2009; Rayson, 2008), starting from keyword extraction and classification of the different collocation patterns for each separate subcorpus. Next, keyword lists of each subcorpus shall be compared and outcomes contrasted. The frequency of the different collocation patterns shall be mutually compared one against the others (Corpas, Ha, Mitkov, 2008). Finally, outcomes obtained shall help us determine the text terminological coverage. Findings on this comparison can provide us with a fuller picture of how the selected terms work and with useful insights into salient markers of Maritime English language.

References:


(77)

Louw, William (University of Coventry and University of Zimbabwe, Zimbabwe) & Milojkovic, Marija (University of Belgrade, Serbia): Shared Logical Form or Shared Metaphysics? In search of corpus-derived empathy in stylistics

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORAs

The Plato website at Stanford University identifies a large number of potential sources for the aetiology of empathy (Greek feeling or suffering together). Apart from ancient history we are given the choice of largely mentalist avenues for pinpointing the source of what is arguably the most important aspect of literary engagement between writer and reader. Within philosophy there is the problem of other minds. Within neurophysiology there are debates on simulation. Hermeneutics singles out the method used within the human sciences. And to the eternal chagrin of Frege moral and scientific psychology take up the balance of a 26 page entry.

However, the use of corpus stylistics is simply not mentioned. But surely it offers very attractive experimental assumptions. If it is common cause that writers communicate shared feelings and suffering to readers and if critics are apt to call these feelings empathy, this ought to be an area the definition of which might be made corpus-attested (Louw 2008). And if a reference corpus were to be used in order to extract situational verisimilitude between texts that might arouse in the reader responses that are akin to suffering or feeling together, we ought to be on a more direct path than those suggested in the Plato collection for running to ground exactly where empathy resides.

Now, of course, corpus stylistics driven as it was in its infancy by human intuition has already explored similarities of vocabulary to the point of diminishing returns if not of nausea. But intuitive opacity has until recently (Louw 2010a, 2010b; Louw and Milojkovic 2014 and forthcoming) prevented the search for empathy through the use of corpus-derived subtext as logical form (Wittgenstein 1929).

Could it be that literary texts that share logical form also share reader-reactions akin to what critics term empathy?

This paper will attempt not only to decide the question posed by its title but also to observe the ways in which the latent largely a priori collocates of the wildcarded grammar strings alter the meaning of the lexical collocates that actually occur within the target text. If this project is successful, the results will be found to be so hugely empirically reliable that the list on the Plato site may find itself appreciably diminished in the interests of science.

The best place to look for a respectable starting point within analogue critical studies and its interface with philosophy would involve the creation of a merger between hermeneutics and corpus stylistics (Teubert 2010; Bleicher 1980). Gadamer (1989) provides an assumption that is readily convertible into an active computational hypothesis at both the surface and the subtextual level. He argues that the significance of a text is not tied to it's author's intentions In writing it. The use of this assumption will reopen the debate in analytic philosophy surrounding the extent to which analysis takes place in a hermetically sealed environment or whether additional information seeps in as part of the process of analysis. The term 'seeps' may need to be reconsidered as reference and authorial corpora often provide relevant additional information with the empiricism of a fire hose. If the empathy of argument is gathered successfully albeit as a by-product of analysis this may go some way to providing subtextually the empathy for which Teubert yearns as he searches for a match in the metaphysics of corpora devoted to societal matters. Literary examples will be provided.
References:


(78)

Lozano, Cristóbal (University of Granada, Spain): Uncovering syntax-discourse factors in L2 Spanish anaphora resolution in the CEDEL2 corpus

PANEL: CORPORAs, LANGUAGE ACQUISITION AND TEACHING

While second language (L2) researchers have traditionally relied on (quasi)experimental data, they have recently started to use learner corpus data (Myles 2005, 2007). Within the framework of interlanguage annotation (ILA) (Díaz-Negrillo & Lozano 2013) and learner corpus research (LCR) (e.g., Granger 2009 inter alia), this presentation shows how corpus data can reveal unexpected L2 behaviour that has gone unnoticed in experimental studies on anaphora resolution.

The bulk of experimental research on L1 English–L2 Spanish reveals a robust pattern (Al-Kasey & Pérez-Leroux 1999, Liceras 1988, Lozano 2002): learners acquire early the fact that overt and null referential pronominal subjects can alternate syntactically, (1). But such (apparently free) alternation is constrained discursively in native Spanish: null pronouns encode topic continuity (Ø in 2), while overt pronouns encode topic-shift when a change of referent is required (él in 3). Importantly, recent experimental L2 studies indicate that learners show persistent deficits at the syntax-discourse interface (Margaza & Bel 2006, Pérez-Leroux & Glass 1999, Rothman 2009): they often produce (i) an overt pronoun in topic-continuity contexts, which causes redundancy (él in 2), and (ii) a null pronoun in topic-shift contexts, which causes ambiguity (Ø in 3), as also reported for L2
Italian (Sorace & Filiaci 2006). L2 Spanish corpus-based studies also point in the same direction (Lozano 2009b, Montrul & Rodríguez-Louro 2006).

(1) Él/Ø es millonario.
‘He/Ø is a millionaire.’

(2) Pedro tiene mucho dinero y #él/Ø dice que #él/Ø es millonario.
‘Pedro has a lot of money and *he/Ø says that *he/Ø is a millionaire.’

(3) María y Pedro viven felices, pero él/#Ø es pobre.
‘María and Pedro live happily, but he/#Ø is poor.’

Building on previous experimental research, a fine-grained ILA scheme (Figure 1) was designed to take into account the multiple factors intervening in anaphora resolution in an L1 English – L2 Spanish learner corpus (Corpus Escrito del Español L2, CEDEL2: Lozano 2009a, Lozano & Mendikoetxea 2013) at upper-advanced proficiency level, as well as an equivalent Spanish native subcorpus. UAM Corpus Tool (O’Donnell 2009) was used to tag and analyse the CEDEL2 corpus, whose data reveal several important findings that have gone unnoticed in previous experimental research. In particular, despite their high level of proficiency:

(i) Learners not only use a redundant overt pronoun to mark topic-continuity, but they also produce full NPs (Figure 2).

(ii) Learners can mark topic-shift via an overt pronoun, as would be predicted for native Spanish, (él in 3), though they drastically prefer using a full NP (Fig. 3).

(iii) Additionally, learners also show a tendency to produce informationally richer phrases than pragmatically required (full NP > overt pronoun in topic-shift contexts; overt pronouns and full NP in topic-continuity contexts), which runs against economy principles (Fig. 4). These deficits have to do with the number of potential antecedents of the anaphor, coupled with the gender distinction of such antecedents.

Corpus data thus reveal that learners prefer being redundant and uneconomical to ambiguous, a finding not previously reported in experimental studies. It will be finally argued that naturalistic learner corpus data can (and should) be used as a follow-up to experimental data to explore new patterns of L2 production (cf. Gilquin 2007).

(79)

Mahloane, Malefu (University of Witwatersrand, Romania): Challenges Encountered in Corpus-based Lexicography for Southern Sotho.

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Natural Language Processing projects for most Bantu Languages of South Africa are at a snail’s pace. Among these resource scarce languages is Sesotho (Southern Sotho), which among other things, does not have an available and accessible corpus, a standard tagset, and a monolingual electronic or print dictionary, etc. This presentation will discuss the challenges encountered in a current project that aims to produce a monolingual electronic Sesotho learners’ dictionary, using a corpus-based approach. These challenges are related to both the compilation of the corpus for Sesotho, using existing language processing tools, as well as challenges encountered with the software employed for the dictionary design, TshwaneLex. Between the 19th century to the 20th and 21st centuries, Sesotho lexicography has not seen significant change in methods, styles and tools used to produce its dictionaries (Kiango, 2000). In South Africa, the language has only produced
multilingual/bilingual dictionaries in both print and electronic forms. As a result, Sesotho remains one of the Bantu Languages that continue to suffer criticism by lexicography researchers, directed to the lack of production of different other types of dictionaries (Nkomo and Wababa, 2013).

The initiator of the project in discussion has found it significant to have a monolingual electronic dictionary for Sesotho. It was found further compelling for this dictionary to be a learners' dictionary. This is as a result of the terminology comprehension and conceptualization factors that affect learners who are not L1 speakers of English, which is a major medium of instruction in South African schools. Sesotho has borrowings from English and some of the borrowed words tent to appear even in Sesotho written texts, without there be Sesotho comprehensible equivalents created to represent them. For example, borrowed words such as demokrasi 'democracy', thekenoloji 'technology', and ithaviu 'interview' were found in Sesotho learners' literature and course books that were used as part of the corpus for the project in discussion. In as much as teachers can be trusted to explain these words to learners in a Sesotho class, they themselves have not a standard reference for definitions (i.e., a dictionary) and chances are that they explain these terms translating from their core meaning in English which in itself may be problematic.

Therefore, the vision for the monolingual dictionary that this project aims to produce is to have explanations of both primary and borrowed Sesotho words in a language which learners conceptualize the world in, and with support from Psycholinguistics, the project initiator attests to a psycholinguistics observation that, once an individual understands concepts in their L1 it is easier to comprehend the concepts when they are expressed L2 (Carroll, 2008). Lexicography, Corpus Linguistics and Terminology theoretical frameworks have been used as guidelines for this project. Additionally, an empirical study was conducted with both high school learners and teachers from whose school books some of the corpus was sampled. Although the project will ultimately add onto lexicography and corpus development for Sesotho as well as indicating the usefulness of the two in language preservation and learning, it is hoped that the challenges that are presented will be addressed by developers of language processing tools to accommodate this language and other resource scarce languages.

(80)

Marcos Miguel, Nausica (Universidad Europea de Madrid): Textbook Consumption in the Classroom: Analyzing Classroom Corpora

PANEL: CORPOR A, LANGUAGE ACQUISITION AND TEACHING

Textbooks permeate foreign language teaching: they structure the lesson and often times serve as the leading curriculum (e.g., Harwood, 2014; Hutchinson & Torres, 1994; Shawer, 2010). Moreover, teachers adapt their textbooks according to their specific context taking into consideration issues such as learners' needs, teachers' preferences, or classroom time (see Tomlinson, 2012). For classifying textbook's adaptations, McDonough and Shaw (2013) have proposed five categories: adding, deleting, modifying, simplifying, and reordering. Nevertheless, these criteria have not been yet subject to empirical examination. Thus, the question remains how teachers utilize their textbooks within a lesson, how this usage shapes their curriculum, and whether adaptations increase the communicative value of modified activities.

This study analyses textbook’s use in a Spanish L2 multi-section course at a large North American University where Teaching Assistants (TAs) were the main group of language instructors. In this instructional setting, textbooks are “a source for in-class activities” and
"a tool for vocabulary and grammar acquisition" (Willis Allen, 2008, p. 11-14). TAs carry out their graduate studies while simultaneously teaching and being trained to teach. In that context, knowing the kind of adaptations in place can contribute to TAs’ pedagogical training and help them to optimize textbooks’ use. Adaptations and variation among instructors should not be negatively seen, though. They can showcase the TA as a critical consumer who adapts the textbook because of on-going classroom needs.

Three TAs, teaching different sections of the same intermediate level course, were observed and audiotaped during the teaching of a book chapter, i.e., five lessons of fifty minutes each. The lessons were verbatim transcribed. Two systems were used to code those events where teachers utilized the textbook in their lessons. On the one hand, Shawer’s (2010) classification of curriculum transmitter and curriculum developer’ strategies were used to analyze the main use of the textbook. On the other hand, McDonough and Shaw’s (2003) criteria were used to identify possible adaptations of each textbook.

Preliminary results show that these teachers were in a continuum between curriculum transmitters and curriculum developers. Adaptation of activities was a very frequent process showing instances of all McDonough and Shaw’s (2003) categories. Despite using the same textbook and syllabus, the lessons of each teacher turned out to be very different because of their adaptations. These TAs tended to increase the discussion around cultural and literary topics which increased the communicative value of each activity.

This study contributes to current methodology on research on foreign language textbooks’ consumption (see Harwood, 2014) by utilizing a classroom corpus. A methodological discussion on the role of classroom corpora for foreign language textbook research, such as the Flensburg English Classroom Corpus (Jäkel, 2010) or the Multimedia Adult English Learner Corpus (see Reder, Harris & Setzler, 2003), will conclude this presentation.

References
(81)

Marín Pérez, María José & Fernández Toledo, Piedad (University of Murcia, Spain):
The influence of cognate meaning detection on the acquisition of legal terminology: a help or a hindrance? a corpus-based study

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

This paper presents a corpus-based experiment which focuses on the influence of Spanish/English cognate meaning detection on the acquisition of legal terminology.

The use of corpora in language instruction has been profusely reviewed by scholars (Johns 1997; McEnery and Wilson 1996; Sinclair 2003; Hunston 2007; Flowerdew, 2009; Boulton 2012), who underlining the advantages and disadvantages of resorting to them as sources of information for the elaboration of didactic materials. Nevertheless, in spite of the large number of experiments carried out to evaluate the efficacy of DDL methods, very few of them focus on the area of legal English (see Boulton, 2010). In addition, Marín and Rea (2012) highlight the fact that there are very few legal English corpora available online, this is why the British Law Report Corpus (BLaRC) was created by Marín (2014).

Concerning the study of cognates in ELT, the panorama seems particularly complex as regards specialised languages, where general words often specialize into technical terms. Legal English terms with Latin origin might be specifically problematic for Spanish learners and more so in the case of disciplines as old established as Law, with many words of common origin, including false cognates.

In view of this, our study focuses on the semantic equivalences between the general and specialised uses of legal terms and on the degree of equivalence between both meanings in the L1 and the L2. We decided to draw upon the BlaRC corpus (8.85m words), due to its potentiality as a source of authentic specialized discourse in which to find examples of legal cognates, and also referred to LACELL (20m words), a general English corpus owned by the LACELL research group at the English department of the University of Murcia, which was used to obtain the general discourse contexts in which the same terms can be found.

We hypothesized that the closer the general and specialised meanings of terms like conclusion, track, battery or conviction, the less difficulties they would pose for the students’ elicitation of their meaning in context. Other variables such as the students’ language competence level or the presence of the selected terms among the most frequent words of the BNC were also taken into consideration. A sample of 56 participants was singled out among a group of Spanish students of Legal English in the first year of their Law degree. They were asked to translate a set of legal terms, each of them appearing both in a general and a specialised context –extracted from LACELL and BlaRC- respectively. In general terms, the results attested that the greatest difficulties found by the students were caused by terms which acquired a new meaning within the legal context (e.g. battery or conviction). Moreover, as it was expected, the higher the students’ language competence level, the better the results and, consequently, the less influence of the type of context on their performance.

The results suggest a need of raising awareness about the double effect of L1 and general English prior knowledge in cognate meaning detection, and point to a need of further corpus-based research that covers a wider range of disciplinary areas. Corpus-based task design certainly appears as a good source for the overt tackling of false cognates in the ESAP classroom based on authentic, prototypical examples from the own students’ practice realms.
REFERENCES


MARÍN, María José, REA, Camino (2012). “Structure and design of the BLRC: a legal corpus of judicial decisions from the UK”. Journal of English Studies, 10. La Rioja: Servicio de Publicaciones de la Universidad de La Rioja


(82)

Mccafferty, Kevin (University of Bergen, Norway): Ø not saying this before your backs: BE-deletion in Irish English, 1731-1840

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

BE-deletion is a salient and heavily-researched feature of African American English (e.g., Rickford 1998; Kautsch 2002; Green 2002; Weldon 2003). Since this is taken to be a creole or African substrate feature in AAVE, it is assumed that superstrate input Englishes from Britain and Ireland did not include this feature. Certainly, BE-deletion has seldom been noted in Britain and Ireland, though there are reports from present-day northeastern England (cf. Tagliamonte 2012:39) and Scotland (Macaulay 1991), as well as nineteenth-century Yorkshire (Giner & Montgomery 1997). Significantly, Hickey (2007:176-177) notes BE-deletion in a wider set of contexts (not just copula deletion) from southeastern Ireland.

This paper uses the Corpus of Irish English Correspondence (CORIECOR) for a diachronic study of BE-deletion in eighteenth- and nineteenth-century Irish English. The findings indicate that BE-deletion was a widespread phenomenon in Irish English historically: geographically, it is found from (London) Derry and Donegal in the northwest to Waterford, Cork and Limerick in the south.

As new data sources become available for regional varieties of British English, a superstrate source may ultimately be documented. At present, however, the earliest British attestations so far postdate Irish English usage in CORIECOR. Meanwhile, the time-depth of 'be'-deletion in Irish makes the Irish substrate a certain contributor to the existence of BE-deletion in Irish English, which often occurs in similar contexts to copula
absence in Irish. Also, BE-deletion was present in Irish English in time to be carried abroad by Ulster Scots in the eighteenth and other Irish emigrants in the nineteenth century.

This study thus adds historical evidence to support Hickey’s (2007:177) contention that attestations of BE-deletion in Irish English must lead to some revision of accounts assuming that copula deletion does not and did not occur in British and Irish Englishes (e.g., Rickford 1998:187). Ireland supplied large proportions of the English-speaking settlers to North America in the eighteenth and nineteenth centuries. Dialect input from varieties of Irish English must be taken into account when considering copula deletion in North American Englishes and other varieties where Irish English input is likely as a result of immigration and contact between relevant groups.

References


analyst’s intuition would allow, but all this information is still in the text. There can be no additional information not contained in the text. Intertextuality is on the margin of such a line of reasoning because it involves absent texts, but still the common denominator is present in the text under study. The wider context is not in it but it is present as an influence. In short, the analytic philosophers appear to have been right.

Louw (2008) calls for corpus-attested definitions of literary devices, such as irony, metaphor, or antithesis. This noble task, if undertaken and fulfilled, would make all stylistics corpus-attested. However, the scientist would encounter a seemingly insurmountable difficulty: he or she would proceed from the long-standing and widely applied, but still analogue and intuitive, classification. For example, what is metaphor? In other words, how can we produce a corpus-attested definition of a concept whose existence is in itself not confirmed in corpus terms?

Concepts are outside of texts. They are formalised pre-computational attempts at generalisation, in the absence of a better analytic tool. The given does not have a conceptual status, it is assigned one by consent of a group of scholars. If we treat this as attempts at intertextuality, there must be formal criteria determining the classification of usages into categories, such as metaphor. For example, in each case of metaphor, a grammatical string contains a lexical item which

- is unique (not found on the list of the QPVs for this grammar string),
- can be replaced by a QPV from the reference corpus, and
- interacts with context clues, in collaboration with which a third lexical item is inferred.

This third lexical item emerges as the target of the metaphor and, ideally, together with context clues, will create states of affairs in the reference corpus that will act as scientific proof that this is indeed the target of the metaphor.

A published example exists in Louw and Milojkovic (2014). In the poem ‘The Circus Animals’ Desertion’, the persona says: ‘Now that my ladder’s gone...’ The context clues suggest that the poet’s ‘ladder’ is his ability to find a ‘theme’, or his ‘circus animals’ from Stanza 1 (another metaphor (what is metaphor?) which we know by analogy (what is analogy?) with the job of a ring master). The subtext of the grammatical string ‘now that my * is’ says ‘a very significant person; an essential quality without which life or honourable existence is impossible’. Therefore, the ‘ladder’ is the persona’s inspiration (essential quality, connected to the ‘circus animals’), but it could be Maud Gonne herself (a significant person; collocating with ‘gone’; the subtext of the first string in the middle section is multiple and includes the theme of despairing of love and asserting love; there are other context clues). In short, the target of the metaphor is inspiration and, on the level of subtext, there are clues pointing to it being a loved person.

One should beware of pre-computational distinctions directing one’s research, as they would make it unfalsifiable in Popperian terms (McEnery and Hardie 2012). With analogue distinctions acting as a springboard but not as formal directions, we must carefully construct the reality of discourse, and may have to call it different names.

Unpublished examples will be provided.

References:
The first step is to replicate Google's experiment using the Twitter messages in Spanish data by expanding rumors and trending topics. In our language corpora and to the algorithm dynamics, which pollute and manipulate attribute those errors to 'big data' overestimation (versus the 'small data' that we can find in our language corpora) and to the algorithm dynamics, which pollute and manipulate data by expanding rumors and trending topics.

The first step is to replicate Google's experiment using the Twitter messages in Spanish.

(84)
Moreno-Ortiz, Antonio & Fernández-Cruz, Javier (Universidad de Málaga, Spain): Identifying Polarity in Financial Texts for Sentiment Analysis: a corpus-based approach

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

In recent years, sentiment analysis or opinion mining has become an increasingly relevant sub-field within text analytics that deals with the computational treatment of opinion and subjectivity in texts. Most Sentiment Analysis systems have focused on specialized domains using domain-specific corpora as training data for machine learning algorithms that classify an input text as either positive or negative. Other systems are lexicon-based, where sentiment-bearing words and phrases are collected and then searched for during analysis to come up with a certain sentiment index. In this paper we describe our methodology to integrate domain-specific sentiment analysis in a lexicon based system initially designed for general language texts. Our system has shown reasonably good results across different types of texts, but falls short as the specialization level increases, since sentiment is lexicalized differently to some extent. Our approach to dealing with specialized domains is based on the idea of "plug-in" lexical resources which can be applied on demand. In order to acquire such resources we employ a simple 3-step model based on the weirdness ratio measure to extract candidate terms from specialized corpora, which are then matched against our existing general-language polarity database to obtain sentiment-bearing words whose polarity is domain-specific.

(85)
Moreno-Sandoval, Antonio (Universidad Autónoma, Madrid) & Moro, Esteban (Universidad Carlos III, Madrid): «Big data» versus «small data»: the case of "gripe" (flu) in Spanish

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

This paper’s main objective is to explore the following statement from Lazer et al 2014 with texts in Spanish:

"'Big data hubris' is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis."

Lazer et al 2014 analyze Google Flu Tool's (GPT) failure to predict accurately the flu season. If in the first 2009 version of GPT, "big data were overfitting the small number of cases" and "GFT was part flu detector, part winter detector", the new GFT version "has been persistently overestimating flu prevalence for a much longer time". These authors attribute those errors to 'big data' overestimation (versus the 'small data' that we can find in our language corpora) and to the algorithm dynamics, which pollute and manipulate data by expanding rumors and trending topics.

The first step is to replicate Google's experiment using the Twitter messages in Spanish.
that were geolocalized and included the word ‘gripe’ (we will call this ‘our Flu Corpus on Twitter’). The hypothesis is that the increased number of messages with that word is a predictor of an approaching peak of cases. The way to verify this prediction is to check the reported cases in the Spanish Health System, which offers the real data sent by physicians in health centers and hospitals. To avoid noise caused by institutional or press messages ("100.000 personas aún no han pedido vacunarse contra la gripe"), we have eliminated all the messages that contain a URL. The results show that the messages on Twitter in Spanish also magnify the real cases of flu (see Figure A), as GPT does (http://www.google.org/flutrends/es/#ES).

The next step is to analyze the Flu Corpus on Twitter in detail to discover what factors contribute negatively to the prediction. For example, the cases in which figurative and humorous language is used: “No es nada, no sé si es un constipado virulento o una gripe virurápida.” "Malo no lo siguiente .. En modo constipado casi gripe ...." Examples like these show that the difficulty in separating the real cases from the figurative ones requires a good management of intentional and pragmatic aspects. However, sometimes there is no irony, the only thing that happens is that the text mentions ‘gripe’, but it does not imply that the subject suffers from it: “Acabo de ver un anuncio de Glocatil gripe y me ha acordado de @...”

In order to refine the results of the simple search for the keyword ‘gripe’, we have tried more traditional methods from Corpus Linguistics: finding synonyms and variants of ‘gripe’ in the Twitter corpus (‘gripazo’, ‘griposo’, as well as segments that indicate that someone has contracted this disease: “con la gripe en casa” “menudo gripazo he pillado” “tos y moqueando” “la gripe me mata”). Discarding orthotypographic variations, there are very few lexical and syntactic patterns in the messages on Twitter.

By contrast, we analyzed the word ‘gripe’ in a medical corpus, Multimédica (Moreno y Campillos 2013) using the tool The Sketch Engine (Kilgarriff et al 2014) – just 319 occurrences against 2764 in Twitter-. ‘Gripe’ usually collocates with ‘virus’, ‘brote’, ‘caso’, ‘estación’, ‘azote’, ‘epidemia’, ‘vacuna’ or ‘temporada’. These collocations counterbalance the ‘scarcity’ of information in the Flu Corpus on Twitter.

In conclusion, our data support the hypothesis of Lazer et al. 2014 that states that “instead of focusing on a ‘big data revolution,’ perhaps it is time we were focused on an ‘all data revolution,’ where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.” In terms of corpora, small but well selected collections of linguistic data should be combined with large repositories from internet and social networks, since sometimes ‘small data’ offer information that is not inferred from ‘big data’.

References:
Google Flu Trends: http://www.google.org/flutrends/es/#ES
Mujcinovic, Sonja (University of Valladolid, Spain): The analysis of subjects in the oral and written production of L2 English learners: transfer and language typology

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

This study considers the oral and written production of English learners and focuses on the analysis of sentential subjects and the issue of transfer.

Previous studies on the acquisition of second languages (L2) have argued that typologically different languages often influence each other resulting in negative transfer (e.g. Odlin 1989, Meisel 2001, Pladevall Ballester 2012). In this respect, transfer is said (1) to have a specific directionality, since structures are often transferred from the first language (L1) into the L2 and (2) to be linked to the amount of exposure. However if the languages exhibit similar grammatical properties, no transfer is expected in this particular area of grammar and amount of exposure should, therefore, play no role.

In the case of sentential subjects, most works on L2 acquisition deal with transfer between two typologically different languages (e.g. Brice & Rivero 1996, Gottardo et al. 2001, Gebauer et al. 2013). In the case of L1 Spanish – L2 English, an over production of (illicit) null subjects in the L2 has been shown to occur and it has been attributed to the influence of the L1 where null subjects are a "legal" option (e.g. Montrul & Rodríguez Louro 2006, Montrul 2010). However, not so much has been said about typologically similar languages (e.g. de Prada 2009, Filiaci 2010). The present study addresses this last issue and deals with the similarities that English and Danish have regarding sentential subjects and how these similarities can be an important factor when dealing with transfer in oral and written tasks in the L2 English production of L1 Danish children. Given that both languages require their subjects to be overt, the production of null subjects is not expected to occur except for the grammatically adequate ones (for example when dealing with coordination). Also, this lack of non-native-like subjects is not supposed to be correlated with exposure so that, regardless of the degree of proficiency in the L2, no such subjects should occur.

In order to test these hypotheses and to deal with (the lack of) transfer in typologically close languages, this study analyzes the English subjects produced by 20 primary school students whose L1 is Danish. Participants were divided into two proficiency groups depending on the amount of exposure to English at school (2 or 4 years). To obtain the written data the participants have been asked to create a story from a set of 5 pictures adapted from The Edmonton Narrative Norms Instrument (ENNI) (Schneide, Dubé & Hayward 2005). The oral data were obtained through a semi-guided individual interview which was audio recorded and then transcribed. Subjects produced in both tasks were classified using 3 criteria: form (full DPs, pronouns or null subjects); grammaticality (correct or incorrect); and appropriateness in terms of referentiality (DPs for referent introduction, disambiguation or emphasis and pronouns for referent maintenance).

The results show that the subjects produced by these English learners are both grammatically correct and pragmatically adequate. They also show that this pattern does not change between the two age groups. Therefore, the analysis of transfer should look into language typology as a primary source for transfer, rather than into amount of exposure.
**PANEL: CORPUS AND LINGUISTIC VARIATION**

This article analyses the linguistic choices made in a corpus of brand slogans. Our main hypothesis is that these choices are determined by the socio-semantics of two factors: repetition and variation. Repetition and variation are twin strategies to enhance memorability by impacting on the psychology of the potential consumer. Both rhetorical strategies are necessary and will sometimes overlap as the product or service being promoted is made known to the consumer.

This paper describes a project where we designed and developed a corpus of brand slogans. This research is based on the collection and subsequent linguistic analysis of brand slogans registered in 2011 in the US Patent Office (USPTO). Currently, there is no clearly defined or exhaustive classification of linguistic resources used in the creation of brand slogans. Most studies are based on particular linguistic aspects such as suffixes (Stvan, 2006), phonetic effects (Yorkston & Menon, 2004), the impact of the use of metaphors (Noble et al., 2013) and the interpersonal aspects of brands (Delin, 2005). Our study has been carried out over a three year period and represents a much wider research project.

The result of our research is a classification system at four levels: phonological level, lexico-grammatical level, syntactic level, semantic level. Within each level, subsections have been created and we can highlight the following rhetorical resources on the phonological level: alliteration, rhyme, rhythm, homophones; on the lexico-grammatical level: the formation of new words and expressions, intertextuality and the use of foreign words; on the syntactic level: parallelism and enumeration; on the semantic level: metaphor, simile, personification and antithesis.

The paper gives an overview of the four levels where extensive analysis was carried out. In our presentation of examples at all four linguistic levels, we concentrate on the concepts of repetition and variation to illustrate these two fundamental linguistic choices and how they are realized in this genre.


PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The issue of the polysemous word sense modelling, together with searching for distinctive patterns of word use in texts and describing possible ways of making sense, continues to attract the attention of modern linguists and NLP systems developers.

The paper looks into the problem of the corpus-based automatic analysis of word contexts in direct and figurative senses in order to construct generalized models of polysemous words used in direct and figurative senses. The main method implies corpus study of the distributive contextual model of polysemous units. We select typical components of contexts for literal and figurative senses of nouns (typical predicates and modifiers, different classes of associated words). Grammatically agreeing modifiers (adjectives and participles as premodifiers) and non-agreeing modifiers (postmodifiers in the Genitive case) are of particular interest for us since such collocations facilitate identification of semantic shifts and description of mechanisms of new senses construction in the text.

Working on a large volume of experimental corpus data for languages of differing structures (Russian and Tatar) enabled us to reveal universal and language specific mechanisms of new senses construction. We used data from National Corpus of Russian Language (www.ruscorpora.ru) and "Tugan Tel" Tatar National Corpus (http://webcorpora.net/TatarCorpus/search/?interface_language=ru). The data from these corpora gave us an opportunity to study statistical sense distribution of polysemous words, to improve existing classifications, to explore the context models of lexical polysemy, and to identify different types of collocations, as well as to resolve their ambiguity.

The study was conducted on parallel Russian-Tatar selection of words belonging to the lexical class of names of natural phenomena, including phenomena and objects of natural origin, names of animals and plants, parts of landscape, etc. Words denoting natural phenomena are characterized by semantic multidimensionality, which is ontologically (subjectival correlation of names) and cognitively (various parameters of categorization and discretization of the meaning) conditioned, as well as by distinct language specificity. Names of natural phenomena easily acquire diverse metaphorical senses that are often fixed in the system of language and are expressed by various collocations.

Typical contextual components (lexical, syntactic), associated with direct senses and the types of figurative senses are identified in the structure of generalized contexts. Thus, the mechanisms of recognition of new senses should be based on the analysis of various components of generalized contexts.

We implemented semi-automatic processing of corpus data by means of specialized tools of "OntoIntegrator" system. We created specialized software that assesses the syntactic resemblance of contexts for each sense of a polysemous noun, collects statistical information on contexts composition, obtains and statistically evaluates collocations containing polysemous words. We chose the set of contextual features and evaluated its weights for the experimental sample for each target noun. As a result of the study we compiled a list of lexical and grammatical characteristics that enable contextual word sense disambiguation of polysemous words from the experimental sample.
References

(89)
Noguera, Yolanda (Universidad Politécnica de Cartagena, Spain): A subfield of English for Submariners

PANEL: CORPUS, LANGUAGE ACQUISITION AND TEACHING
For the purposes of this paper, I reach definitions and framework to promote the study of Submarine English through a small corpus pilot study. The selection and grading of items for the syllabus of an English for specific purposes book, mainly the hapaxes and the authentic contextualization of some specific nouns worth considering. All these contents will be analysed using the Wordsmith tools, establishing a lexical approach for our pedagogical purposes. Findings related to a sub-field of Submarine English concerned with "Salvage and Rescue" language. To study it, it has been used a NATO unclassified book entitled “ATP-57” which is used as the compulsory textbook by teachers and students to study “Salvage and Rescue”.

(90)
O’Donnell, Mick (Universidad Autónoma de Madrid, Spain): Exploring the use of quantifiers in Spanish learners of English

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING
This paper will present a study of how nominal quantifiers are used by University-level learners of English in Spain. The study is based on the 540,000 word WriCLE corpus (Rollinson and Mendikoetxea, 2010), a set of 560 short essays from students in an English Studies degree, with each essay associated with a CEFR proficiency score.

A program was developed to automatically re-process the output of the Stanford parser (Klein and Manning, 2003), to recognise each noun phrase, identifying the occurrence of any quantifier (e.g., all, none, some, many, much, lots of, little, few, etc.). The context of occurrence of the quantifier is also featured, recording the syntactic position (e.g., predeterminer, determiner, post-determiner), the nature of the head noun (count or mass,
singular or plural), as well as the overall context of use, e.g.,
   a) imperative, declarative or interrogative Mood;
   b) positive or negative polarity;
   c) presence of intensification (e.g., too much) or comparison (e.g., so much);

The goal of the research is to identify the changing patterns of use of these quantifiers from the A2 level (the lowest in the corpus) up to the C2 level, identifying changing degree of use with rising proficiency. A comparison with native use will also be made, using a subsection of the BAWE corpus (Nesi et al, 2005), which has also been parsed.

We will also report on the kinds of errors the learners are making in the use of these quantifiers, e.g., use of ‘much’ with a count noun, the use of ‘any’ in a positive statement, etc. Our post-processor has been programmed to identify incorrect uses of quantifiers, at least as far as lexical and syntactic information permits.

The identification of over-use and under-use of particular quantifier structures provides valuable input to the EFL curriculum, as does the identification of the highest frequency quantifier errors at each proficiency level.

References


(91)

Oncins, Jose Luis (Universidad de Extremadura, España): Anglicismos idiomáticos en el español actual: el caso de "darse un tiro en el pie" (<"shoot onself in the foot")

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

Entre los muchos anglicismos que el español recibe a diario del inglés encontramos un buen número de expresiones idiomáticas, en su forma original o, más frecuentemente, traducidas como calcos. Éste es el caso de "patata caliente", "salir del armario" o, más recientemente, "darse un tiro en el pie". Muchas de estas expresiones acaban incorporándose al acervo idiomático de la lengua y la cultura receptoras tras un periodo de aclimatación, que será más o menos largo dependiendo, entre otros factores, del grado de opacidad/transparencia semántica de cada expresión. En este presentación se analiza una amplia muestra de la expresión que le da título (casi 500 ejemplos), obtenida de internet a lo largo de dos años y medio mediante la herramienta "google alerts".
(92)

Ooi, Vincent B. Y. (National University of Singapore, Singapore): Examining the GloWbE corpus as a lexicographic resource for Singapore, Malaysian and Hong Kong English

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

With the advent of big data for linguistic research, the GloWbE corpus promises to ‘expand horizons in the study of World Englishes’ (Davies and Fuchs, 2015 – forthcoming). In turn, would such a corpus offer a reliable evidence base for the incorporation of World Englishes – highlighting the pluricentric nature of English as the leading global language – in the dictionary? This paper offers a modest, exploratory answer to such a question by considering certain phrases important in the contexts of Singapore, Malaysia and Hong Kong - three ESL countries which have a number of striking similarities (and differences).

First, GloWbE is said to be based on “1.9 billion words in 1.8 million web pages from 20 different English-speaking countries. Approximately 60 percent of the corpus comes from informal blogs, and the rest from a wide range of other genres and text types.” (Davies and Fuchs 2015). The general idea for GloWbE, as it is for other web corpora, is to regard the ‘web as corpus’ (Kilgarriff and Grefenstette 2003). But, Sinclair (2004), while welcoming the WWW as ‘a remarkable new resource for any worker in language’, also notably warns that ‘the WWW is not a corpus’, if the latter is defined to be maximally representative of the linguistic phenomenon in question. Sinclair’s reasons include the ‘mysterious’ dimensions of the Web and the varying algorithms afforded by the various search engines that do not lead to the right balance and sampling (even if size is exponentially increased).

Examining the GloWbE corpus, the Standard Singapore English phrase “killer litter” is significant in Singapore which is sensitive to the danger/injury posed by heavy objects thrown from high-rise buildings. In the GloWbE corpus, the phrase is remarkably absent in 18 countries and occurs a total of 7 times in Singapore and 1 time in the Sri Lankan context. The one time that it does occur in the Sri Lankan context, as in any given context, is not significant – given the odd migration and diffusion of English use across contexts.

In the Malaysian context, a prototypical ‘Manglish’ (or colloquial Malaysian English) phrase is "lepak" (meaning ‘to skive’ or ‘to chill out’), borrowed from Malay: A search for the phrase in GloWbE yields a view of frequencies from different countries: U.S.A. (1 occurrence), Canada (2), UK (2), Australia (6), Singapore (5), Malaysia (23). In this case, ‘the ability to see the frequency of any word, phrase, or grammatical construction in each of the 20 different countries’ (Davies, 2013) may lead to the misleading conclusion that this Manglish phrase is most used in Malaysia, and then productively more used in Australia than in Singapore. In the Malaysian concordance for "lepak", the ‘chill out’ sense is retained but a closer examination of the Australian concordance shows that it refers to someone’s name (‘Dennis Lepak’) which has no bearing to the Manglish phrase.

Turning to Hong Kong English, "shroff" is a term that many Hongkongers regard as ‘standard English’ because of its prevalence in parking lots/carparks ("shroff" means ‘an office or kiosk, e.g. in car parks’ (Bolton 2003: 295). A search for the term gives the impression that it is used across a number of countries: U.S. (4), Canada (3), UK (12), Australia (4), New Zealand (20), India (126), Sri Lanka (15), Pakistan (31), Hong Kong (16). But, unlike the case of "lepak" in which a quick distinction between upper and lower case would do, the lexicographer will have to trawl through the ‘parking lot’ sense from the proper name sense in the concordance listing for Hong Kong.

In this paper, a number of other linguistic examples will be used to show the internal diglossic nature of these varieties of English (e.g. between ‘standard Singapore English’ and ‘Singlish’) that lexicographers will have to sift through the data resource afforded by
the GloWbe corpus. Notwithstanding this, it would be very much apparent that the GloWbe corpus is a much welcome resource for the lexicographer to incorporate World Englishes into the dictionary.

Select References:
Davies M, and R Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English corpus (GloWbe), In English World-Wide 36:1. (forthcoming), pp1-29.

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

This contribution focuses on the analysis of the interplay of voice-overs (VOs) and on-screen texts (OSTs) in the discourse of television advertising. To our knowledge, in spite of all the work published to date (see Brierley, 2000; Byrne, 1992; Coltrane & Adams, 1997; Fuertes-Olivera et al., 2001; Katz, 2010; Myers, 1994; Pennock-Speck-Del Saz, 2013; Piller, 2006; Schmidt et al., 1995; etc.), this type of research has not been done since Leech published his work English in Advertising: A Linguistic Study of Advertising in Great Britain (1966). In all probability, the reason is that TV commercials are fairly complex discourse products and, consequently, researchers have decided to deal with very aspects they deem to be relevant. However, we believe that examining how VOs and OSTs complement each other is essential to understand how messages are conveyed in TV ads. To carry out this study we used the Multimodal Analysis of TV Ads corpus (henceforth MATVA), which contains transcriptions of British TV ads recorded over six days during the years 2009, 2010 and 2011. MATVA has a total of 2,140 ads. MATVA is one of the largest TV ad databases of its kind in terms of the number of ads. After a pruning process that involved eliminating duplicate ads, we were left with 1122 commercials (see further details in Pennock-Speck & Fuster-Márquez, 2014). While all ads have OSTs, not all of them contain VOs. As we are interested in the interplay of VOs and OSTs, we selected ads containing both types. We thus ended up with a final tally of 785 commercials. Among the strategies we have found the following are prevalent: (1) OSTs ('The new Toyota Avensis sponsors ITV Mystery Drama') which reinforce the message of the VO by repeating it verbatim; (2) OSTs (Thirst Pockets. The power of an elephant in just one sheet) which summarise what is said in the VO (New Thirst Pockets. So absorbent and strong, just one sheet could do the job. New Thirst pockets. The power of an elephant in just one sheet) and (3) OSTs (On DVD now. www.tesco.com TESCO, every little helps Only £12.47) which supplement the VOs with practical information such as the price of the product and websites where further information may be obtained (The official workout. Staring trainers and contestants from the hit ITV show. The Biggest Loser workout, on DVD now). Our talk finalises with a discussion on the discourse-pragmatic strategies found in VOs and
OSTs. Our research indicates that the former are normally made up of texts of a persuasive nature (the top keyword for VOs is "you" which is rarely found in OSTs) while the latter generally include varied product information, information of a quasi legal kind (key words such as "subject", "availability", "conditions") and contact information –the top four keywords for OSTs being "www", "uk", "co" and "com".

(94)
Perea, María Pilar (Universitat de Barcelona, Spain): El contacto lingüístico en un corpus de correspondencia que abarca los siglos xix y xx

PANEL: CORPUS AND LINGUISTIC VARIATION

El CD-ROM “Epistolari d’Antoni M. Alcover (1880-1931)”, publicado en 2008, contiene la transcripción completa de la correspondencia que el lexicógrafo mallorquín recibió entre 1880 y 1931. Alcover conservó 16.005 documentos, que incluyen cartas, postales, tarjetas de visita y otros materiales escritos, 638 de los cuales fueron redactados por él mismo (borradores, copias o cartas que quizá nunca llegó a enviar). Aunque la correspondencia está redactada en catalán, castellano, alemán, francés, italiano, inglés, latín y esperanto, el catalán domina numéricamente con 10.456 documentos (65,3 %). Además de su interés intrínseco, este corpus ha sido utilizado hasta ahora para estudiar la interferencia lingüística entre el catalán y el castellano y también para detectar diversos rasgos dialectales en la morfología verbal catalana de principios del siglo XX.

Esta presentación tiene el objetivo de mostrar cómo el contenido del CD-ROM puede contribuir a analizar la variación lingüística partiendo del corpus formado por los más de cinco millones de palabras que conforman la correspondencia (3.529.159 de las cuales son distintas). En este caso, se presta una atención especial al léxico, en el ámbito del contacto lingüístico, con relación a los préstamos, a la alternancia en código o a otras características propias de la interferencia lingüística que aparecen durante este periodo de cincuenta años, especialmente entre el castellano y el catalán (fallo, barato, etc.), y el francés y el catalán (comité, turisme, etc.), y, secundariamente, entre el italiano y el catalán (adagi, aquarel·la, etc.), y entre el inglés y el catalán (trolei, túnel, etc.). En un primer estadio, las palabras se clasifican en tipologías, y posteriormente se analizan los procesos de importación, adaptación gráfica y transferencia semántica.

(95)
Pérez-Paredes, Pascual, Aguado Jiménez, Pilar & Sánchez Hernández, Purificación (Universidad de Murcia, Spain): Migrants in administrative language in the UK

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

Immigration is at the heart of societal changes these days, although our current migration policies have been enacted for almost a century now. In fact, European migration started a "new epoch" in 1918 when "entry was no longer free and unimpeded" (Isaac, 2013: xi). In 2007, Spain reached an immigration rate of 10 % of the population. In the UK, this rate was 8.1 % that year and had reached its peak in 2005 with some 320,000 Non-EU citizens getting to the UK, according to the Office for National Statistics [2]. Although migration covers both emigration and immigration (Isaac, 2013:4), it is the latter that is regulated by receiving countries.

Fernández Vitories (2013: 64) contends that the EU countries use language both as an
entry barrier and an integration tool. Sancho Pascual (2013:6) maintains that the existence of migrations and multicultural scenarios call for a redefinition of the identity markers of those involved in these processes. In this context, how is the immigrant represented by the administration? How does the receiving country construct the identity of immigrants?

The use of corpora has attracted the attention of researchers in the field of discourse analysis for a myriad of reasons. The exploration of attested uses of language, combined with the computational power and flexibility of existing software, have contributed enormously to the spread of this approach. In this vein, Some researchers have looked at how the identity of minority groups such as gay men (Baker, 2005) or Muslims (Baker, Gabrielatos, & McEnerey, 2013) have been represented in the media. We have relied on the research methodology in Baker et al. (2008) and Baker, Gabrielatos, & McEnerey (2013) for the combination of Critical Discourse Analysis (CDA) and Corpus Linguistics (CL). Our paper examines the uses of the lemma “migrant” in two corpora of immigration law (1.2 M tokens) and information texts (2.3 M tokens) produced by the UK immigration Authorities during the 2007-2011 five year period.

Our results reveal that the term “migrant” is used most often in its singular form in both corpora, mainly pre-modified by noun phrases that typify their adscription to five existing Visas in the context of the UK immigration Tier System, and are predominantly seen as applicants of leaves to remain in the UK. However, there are important differences between the co-texts in both corpora. While in the Administrative Law corpus migrants are mainly represented as fee providers and are subject to heavy demands from the UK administration, this very same body represents them as either highly-skilled or high-value persons in 22% of the uses in the corpus of information texts produced by the UK immigration Authorities. This paper will discuss a detailed analysis of the frequencies and collocational behaviour of the term in both corpora, and will offer insights into the representational strategies most frequently used in both corpora.

References


Sancho Pascual, M. 2013. Dimensión lingüística de las migraciones internacionales. Lengua y migración, 5:2, 5-10.


(96)

Periñán-Pascual, Carlos & Mestre-Mestre, Eva M. (Universitat Politècnica de València, Spain): DEXTER: Automatic extraction of domain-specific glossaries for language teaching

PANEL: CORPORAS, LANGUAGE ACQUISITION AND TEACHING

The use of corpora in terminography is currently a requirement for language teaching. Often, for language teachers, identifying those lexical units which belong to a given specific domain is a complex task, where simple introspection or concordance analysis does not really become effective. For instance, applying standard frequency criteria to a corpus tends to extract general-purpose vocabulary and is therefore of limited use in identifying technical words. Today, there is a variety of open-source corpus analysis software, e.g. IMS Open Corpus Workbench [1], PhiloLogic [2], Poliqarp [3] or XAIRA [4], among many others. These tools, most of for linguistic or lexicographic research, usually integrate a set of utilities which enable users to check word frequency, concordances and collocations.

The use of such tools is becoming crucial for specific language teaching at University level. The goal of this paper is to describe the design of DEXTER (Discovering and EXtracting TERminology), an open-access platform for data mining and terminology management, whose aim is not only the search, retrieval, exploration and analysis of texts in domain-specific corpora but also the automatic extraction of specialized words from that domain. DEXTER adopts a hybrid approach to term extraction from unstructured data collections, where lexical filters for unithood are applied together with a set of termhood statistical measures, to validate candidates on the basis of stemmed ngrams. The modular architecture of this terminology workbench facilitates the processing of corpora in any language and about any specialized domain. This workbench falls within the framework of automatic term extraction, which is currently a priority field of research in the language industries, which enables teachers to create their own glossaries, appropriate to their particular teaching needs. In this regard, its benefits are immediate, particularly in areas such as document categorization, machine translation or ontology development. Also, DEXTER can contribute to FungramKb [6,7], helping improve specialized language processing. Among the many potential applications of DEXTER, this paper focuses on the automatic extraction of domain-specific glossaries for Language-for-Specific-Purpose courses, where scientific and technical documentation is often used instead of a textbook. Because separating specialized words from the general-purpose vocabulary is a labor-intensive and time-consuming task, DEXTER can become very useful for all those teachers who intend to design their own LSP courses.

REFERENCES:

[2] sites.google.com/site/philologic3/home
Taking advantage of the immense possibilities of the web and through the Hipertext Transfer Protocol (HTTP), readers of digital journals and newspapers participate in what has often been termed as a written conversation (Wildner-Bazzett, 2005) to comment, reply, applaud or criticize what has been written and made available through the web. In this paper I have taken as the main subject of discussion a recent crisis, the Ebola outbreak in Spain, and how the international readership has responded to it in terms of what it is, how it is transmitted, how can it be stopped and whether or not Spain has put the necessary measures to control it. The criticism has been abundant and the analysis centers on how participants select their discourse according to their intended purpose in commenting the crisis.

Texts have been drawn from four British Medical Journal editorials and two reports from each of the following two newspapers, The Guardian and The Washington Post. They have been studied in terms of the linguistic features in both disciplines and in the responses from the readership to each discipline, academic editorials and newspaper reports. The analysis will include the rapid responses to the academic editorials and readers’ comments to the newspaper reports. We understand rapid responses as quick comments to published editorials, medical cases or journal features, allowing for a variety of interlocutors and opinions through different types of discourse (Piqué-Angordans et al., 2010). The main difference with readers’ comments is that rapid responses’ identities of respondents are included; these respondents are usually professionals in the area of health who basically, although not all of them, maintain an academic register. The authors of the readers’ comments, however, by assigning to their messages a nickname, ordinarily remain unknown; thus, by safeguarding their anonymity, they can unduly express their opinions.

Among the main results, the formal vs. informal register is emphasized, in addition to the fact that the rapid responses to the journal articles generally maintain the formal register shown in the source text, while the readers’ comments to the newspaper reports often resemble a cell phone text message exchange (WhatsApp) or entries in an online forum, thus maintaining a plain colloquial register. In these readers’ comments, one can find all sorts of texts, from didactic comments on the crisis at hand or on how the virus can be transmitted, to colloquial expressions, ironic and sarcastic puns, disrespectful discourse, and often simple derisory comments.


Piqué-Noguera, Carmen (Universidad de Valencia, Spain): Linguistic evaluation of readers’ digital feedback to a health crisis: a register approach

PANEL: CORPUS AND LINGUISTIC VARIATION

Although different, register and genre often appear intermingled with each other; in fact, a genre often cuts across several registers (Trosborg, 1997). The difficulty in clearly defining these terms has prompted scholars to try to find answers, such as Biber (1988) proposing his multi-dimensional approach through corpus analysis. Register is connected with the use of the language and varies according to the context of its situation. It also shows the writer’s individuality with the capability of adapting to each situation. This paper approaches the recent Ebola outbreak and how it is rendered through journal editorials and newspaper columns.

Taking advantage of the immense possibilities of the web and through the Hipertext Transfer Protocol (HTTP), readers of digital journals and newspapers participate in what has often been termed as a written conversation (Wildner-Bazzett, 2005) to comment, reply, applaud or criticize what has been written and made available through the web. In this paper I have taken as the main subject of discussion a recent crisis, the Ebola outbreak in Spain, and how the international readership has responded to it in terms of what it is, how it is transmitted, how can it be stopped and whether or not Spain has put the necessary measures to control it. The criticism has been abundant and the analysis centers on how participants select their discourse according to their intended purpose in commenting the crisis.

Texts have been drawn from four British Medical Journal editorials and two reports from each of the following two newspapers, The Guardian and The Washington Post. They have been studied in terms of the linguistic features in both disciplines and in the responses from the readership to each discipline, academic editorials and newspaper reports. The analysis will include the rapid responses to the academic editorials and readers’ comments to the newspaper reports. We understand rapid responses as quick comments to published editorials, medical cases or journal features, allowing for a variety of interlocutors and opinions through different types of discourse (Piqué-Angordans et al., 2010). The main difference with readers’ comments is that rapid responses’ identities of respondents are included; these respondents are usually professionals in the area of health who basically, although not all of them, maintain an academic register. The authors of the readers’ comments, however, by assigning to their messages a nickname, ordinarily remain unknown; thus, by safeguarding their anonymity, they can unduly express their opinions.

Among the main results, the formal vs. informal register is emphasized, in addition to the fact that the rapid responses to the journal articles generally maintain the formal register shown in the source text, while the readers’ comments to the newspaper reports often resemble a cell phone text message exchange (WhatsApp) or entries in an online forum, thus maintaining a plain colloquial register. In these readers’ comments, one can find all sorts of texts, from didactic comments on the crisis at hand or on how the virus can be transmitted, to colloquial expressions, ironic and sarcastic puns, disrespectful discourse, and often simple derisory comments.
References


(98)

Prado Alonso, Carlos (University of Valencia, Spain): A Corpus-based Analysis of And-parenthetical Constructions in British and American English Texts

PANEL: CORPUS AND LINGUISTIC VARIATION

Parenthetical constructions are detached structures (often clauses) which are inserted in the middle of another structure, and which are not fully integrated in the sense that they could be omitted without affecting the rest of the structure.

In the last decade, several types of parenthetical constructions, such as those illustrated in (1)-(4) below, have been the subject of extensive research from a functional perspective (cf. Blakemore 2005, 2006, 2007; or Déhé 2014; Déhé and Kavalova, 2007, among others). This study is a further contribution to this line of research and offers a corpus-based analysis of one type of parenthetical construction namely, and-parenthetical, as shown in (5).

1) They are invited to consider the facts that when a prisoner’s confession, or even his letter home, contained inappropriate words, it was suggest that Chinese People’s Volunteers should be substituted. (Or-parenthetical)

2) In his excellent book Robert Protherough suggest that there is a spectrum between what is objectively correct — that is, something which all speakers of a language will agree on as being ‘there’ in the text — and things which are subjective and purely personal (That is-parenthetical)

3) This predicts that living in an area with people from the same cultural group, as is the case with our Bangladeshi subjects, reduces the risk of mental health problems. (As-parenthetical)

4) I’ve been dreaming for winning a gold medal for what 20 years now. (What-parenthetical)

5) What I’m saying — and I’m really agreeing with Bill here — is that anti-social behaviour orders are the end of the line. (And-parenthetical)

The data for this study are taken from six computerised corpora of British and American Present-day English written texts taken from the BROWN family of corpus: namely the LOB, the Brown, the FLOB, the FROWN, BrE06 and AmE06.
And-parentheticals have been considered speech-bound phenomena (cf. Kalakova 2007) and their analysis has been neglected in the written mode. It is also usually argued that and-parenthetical clauses are the result of a stylistic choice or the result of a sort of on-line reformulation and revision (cf. Blakemore, 2005). Beyond that, however, the analysis of the data retrieved from the corpora will show that these types of constructions are also attested in writing and that they can also be considered markers of persuasion and addressee involvement in discourse. In sum, the paper will shed light on frequency and distribution of and-parenthetical constructions in British and American English Present-day English texts. The data will also show that, in written discourse, and-parentheticals serve to provide the addressee with (background) information expressing the addressee's degrees of commitment, judgements, or opinions in the context of the main utterance.

REFERENCES:
Dehé, Nicole. 2014. Parentheticals in Spoken English: The Syntax-Prosody Relation. Studies in English Language

(99)

Primo Pacheco, Joaquin (Universitat de València, Spain): A Corpus-Based Quantitative Approach to Evaluative Prosody in Literary Discourse

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

The purpose of this paper is to deploy corpus linguistics to quantitatively explore and analyse evaluative prosody (Bednarek, 2006), or the prosodic realisation of appraisal or evaluation (Martin and White, 2005; Martin and Rose, 2007), in a prose fiction text, most importantly in reference to characterisation and the creation of suspense, understood as "an emotional response to narrative fictions" (Carroll 1996: 74).

Overall, appraisal is concerned with interpersonal meanings of language and the "subjective presence of writers (...) in texts" (Martin & White 2005: 1). These interpersonal meanings, in turn, "are often realized not just locally, but tend to sprawl out and colour a passage of discourse, forming a 'prosody' of attitude" (Martin & Rose 2007: 31). As a result, "evaluation extends like a wave over the text and lends a specific 'evaluative prosody' to it" (Bednarek 2006: 8).

In accordance with Bednarek (2006) and the claim made by Martin and White (2005) about the relationship between qualitative and quantitative methodologies in the analysis of evaluation, this contribution follows a CADA (Corpus Assisted Discourse Analysis) approach and advocates that corpus linguistics may be an effective arena with which to inform appraisal analyses with a quantitative approach, along with the qualitative viewpoint that appraisal theory brings per se.

This pilot study focuses on the first chapter of Robert Bloch’s suspense novel “Psycho” (1959), as part of an ongoing research project which intends to analyse from a functional
linguistic perspective the many literary works that suspense film director Alfred Hitchcock adapted. The purpose of this project is to contribute to "the ongoing reevaluation of Hitchcock as an auteur" (Boyd and Barton Palmer 2011: 4) by linking this reevaluation to the element by which Hitchcock is still universally renowned (that of suspense) and in turn, by analysing how suspense is realised and elicited through linguistic means in Hitchcock's original literary sources and in Bloch's "Psycho" for this paper in particular.

To do so, this study draws, on the one hand, on Zillmann’s suspense theory (1996), which emphasises reader alignment and the necessary development of affective and empathetic dispositions towards characters on the part of the reader. On the other hand, it draws on evaluation (Macken-Horarick, 2003; Martin & White, 2005; Bednarek, 2006; Martin & Rose, 2007) to analyse how Bloch makes his readers align or disalign towards his characters through appraisal resources.

On the whole, this contribution looks specifically at the relationship between Norman Bates and his mother, Mrs Bates (named Mother throughout the novel) in the opening chapter of Bloch's "Psycho", the events of which are absent in Hitchcock's 1960 film. This chapter narrates an 'encounter' between both Norman and Mother, who is already dead, a fact which the audience is unaware of at this point. Thus, a qualitative analysis of appraisal resources will demonstrate that readers are positioned towards believing that Norman is a harmless man, whereas Mrs Bates is a tyrannical mother capable of committing crime, a belief which is later questioned and eventually refuted, thus effectively contributing to the overall element of suspense in "Psycho". However, a quantitative corpus-based approach to this qualitative analysis will undoubtedly shed light on the distribution and the frequency of the evaluative nuances that Bloch creates by means of language in order to build up suspense in his novel.

References


(100)

Prinsloo, Danie (University of Pretoria, South Africa): Corpus-based lexicography for under-resourced languages – maximizing the limited corpus

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

The days of a default corpus size of one-million words such as the ground-breaking
BROWN corpus or the LOB corpus being regarded as an acceptable norm are long gone and currently corpora used for the compilation of dictionaries for major languages typically run into hundreds of millions if not billions of words. Balanced and representative corpora reflecting sincere attempts towards corpus designs and representation of stratified speaker groups as well as different levels of corpus annotation and sophisticated corpus manipulation e.g. Sketch Engine, Dante, WordSmith Tools became the norm as an international standard and represent the typical scenario for major languages of the world.

This paper, however, focuses on under-resourced languages for which only very limited corpora are available and how such relatively small and often unbalanced, unannotated raw corpora could be maximally utilized for lexicographic purposes to obtain similar results in the absence of large corpora. African languages, such as the Bantu languages and Afrikaans, will be studied in this regard. The aim is to determine to what extent enlarging a corpus from e.g. one to 10 million, and from 10 million to 100 million tokens enhances its potential for (a) macrostructure compilation, (b) information on the most important microstructural aspects and (c) the creation of lexicographic tools. It will be argued that valuable and even sufficient data for the compilation of a specific dictionary can be extracted from a relatively small corpus of circa one million tokens. The question is how much energy should be invested in the maximum utilization of a limited corpus versus increasing the corpus size and corpus cleaning activities.

On the macrostructural level a qualitative evaluation will be made of a lemmalist compiled from a corpus consisting of one million words versus a 10 million word corpus versus a 100 million word corpus for Afrikaans and a one million word corpus versus a seven million word corpus for Sepedi.

On the microstructural level the evaluation will be focused on the value of information drawn from limited corpora in terms of meaning, sense distinction, examples of usage, collocations and proverbs/idioms.

As for the creation of lexicographic tools it will be shown how even a relatively small corpus of one million words can be utilized to create useful tools such as rulers, block systems, indicators of spreading-across-sources, etc. So, for example, it will be indicated that in the absence of larger corpora a one million word corpus can be sufficient to build a sensible guide for the lexicographer for balancing alphabetical stretches in the dictionary and even that larger corpora do not contribute substantially to the refinement of such tools.

It will be concluded that raw and even unannotated corpora built only from written data, although not reflecting an ideal situation, can substantially assist the lexicographer in the compilation of especially small bilingual and monolingual dictionaries. An attempt towards a schematic illustration of corpus size versus achievement, in terms of macrostructural and microstructural aspects as well as the creation of lexicographic tools will finally be attempted.

References


Dante: http://www.webdante.com/

Sketch Engine: http://www.sketchengine.co.uk/
WordSmith Tools: http://www.lexically.net/wordsmith/index.html

(101)
Rabadán, Rosa (Universidad de León, Spain), Pizarro-Sánchez, Isabel (Universidad de Valladolid, Spain) & Sanjurjo-Gonzále, Hugo (Universidad de León, Spain): GEDIRE: herramienta para la redacción de Directors’ Reports / GEDIRE: A Directors’ Reports Writing Tool

PANEL: SPECIAL USES OF CORPORA

Este trabajo presenta el primer prototipo de una aplicación informática que facilita la escritura del género Directors’ Reports consistente en un generador textual (GEDIRE) desarrollado a partir de los datos lingüísticos extraídos de un corpus especializado (Bowker y Pearson 2002, Hunston 2002).

GEDIRE parte del diseño, compilación y etiquetado de un corpus monolingüe inglés especializado en Directors’ reports (M-En-GEDIRE) compuesto por 120 informes completos (230.646 palabras), escritos en lengua inglesa, publicados en la última década y procedentes de grandes empresas que operan en diversos sectores. El corpus se compilió ad hoc con el objetivo de obtener información lingüística de carácter cuantitativo, que permitiera determinar tanto la estructura propia de este tipo de informes como sus especificidades gramaticales y terminológicas, y se etiquetó por movimientos retóricos (Sarjit Singh et al. 2012, Bhatia 1993, Swales 1990).

Cada uno de los elementos retóricos fue analizado con el fin de describir tamaño en número de palabra y párrafos, tipo de oraciones y conectores, grupo verbal (voz, modalidad, tiempo, aspecto y verbos soporte), fraseología y n-gramas. El resultado de dicho análisis es un complejo listado de datos lingüísticos de carácter recurrente (Scott y Tribble 2006, Oakey 2002, Hunston y Francis 2000) y sin utilidad para un usuario final carente de conocimientos lingüísticos pero que transformados en módulos de líneas modelo de escritura y de glosarios específicos conforman la base de datos a partir de la cual se construye el generador. Los elementos que componen el generador dependen en exclusiva del contenido de esta base de datos, siendo una construcción dinámica; es decir, el contenido de la base de datos puede cambiar sin que esto repercuta en una readaptación o modificación del código fuente del software existente. El generador se creó a partir de los lenguajes de programación HTML 5 (HTML5, 2014) y JavaScript (JavaScript, 2011) con la ayuda de la librería de JavaScript Dojo (Dojo Toolkit 1.6, 2012) para la mejora de los aspectos visuales y transferencia de datos. Alojado en un servidor Web, resulta accesible desde cualquier dispositivo con conexión a Internet a través de un simple navegador.


103
suggests equating the language of learning to the key vocabulary and phrases of the content language of the subject or the specific lesson. There is little doubt that such key vocabulary can be found by analyzing a specific corpus. Therefore, this study suggests equating the language of learning to the key words found by using the keywords tool in Wordsmith (Scott, 1998), the clusters which keywords forms and the significant

PANEL: CORPORAS, LANGUAGE ACQUISITION AND TEACHING

Once again language teaching has found a strategic ally in Corpus Linguistics. This time Corpus Linguistics comes to serving the current approach to learning content subject through the medium of English, the so-called Content and Language Integrated Learning (CLIL). CLIL is adopted as a means to an end, that of acquiring a knowledge and command of at least two foreign languages as promoted by the European Union within a set of proposals for the economic and social fields and for relations with European citizens (European Commission, 2003). CLIL started to be offered at primary and secondary education levels in its mainstream provision in 2004/05 (Eurydice, 2006) and little by little has reached tertiary education. According to Dafouz and Núñez (2009), more than thirty institutions in Spain, at the time of publication, were offering bilingual programs in degrees like Business, Tourism, Law, Telecommunication and Humanities. A bilingual degree in Business Administration was available in 2011 both at the University of Murcia and the Technical University of Cartagena (UPCT), and the current academic year 2014/15 is the onset of a bilingual degree in Telecommunication Engineering at the UPCT. Such upheaval entails renewed teaching methodologies which attach more weight to the vehicular language used to convey content, since language itself is also a learning goal. And here is where TEC (Rea, 2008), a specialized corpus of telecommunication English, comes into play because it embraces precisely the contents of the degree in Telecommunication Engineering in English in addition to other texts coming from its professional realm.

In regard with the 4Cs (content, communication, cognition and culture) conceptual framework in CLIL (Coyle et al., 2010), the authors pinpoint the Language Triptych which consists of the language of, for and through learning that the teacher should consider for a given lesson. The language of learning refers particularly to the key vocabulary and phrases of the content language of the subject or the specific lesson. There is little doubt that such key vocabulary can be found by analyzing a specific corpus. Therefore, this study suggests equating the language of learning to the key words found by using the keywords tool in Wordsmith (Scott, 1998), the clusters which keywords forms and the significant

international.org/ecma-262/5.1/
Oakey, D. 2002. Formulaic language in English academic writing. en E. Reppen et al. (eds.)
Using Corpora to Explore Linguistic Variation, Amsterdam/Phildelphia, John Benjamins,
111-129.
Sarjit Singh et al. 2012. Revisiting Genre Analysis: Applying Vijay Bhatia’s Approach,
Procedia – Social and Behavioral Sciences, 66, 370-379
Cambridge University Press

(102)
Rea Camino (Universidad Politécnica de Cartagena, Spain) & Marín, Mª José
(Universidad de Murcia, Spain): A key perspective on specialized lexis: keywords in
Telecommunication English for CLIL
collocates with which keywords keep company and show their typical use. An added value
is attached to keywords as they manage to fairly reflect the terms in a corpus (Marin and
Rea, 2013; Marin, 2014), so the traditional classification of vocabulary into technical, semi-
technical, academic and general is set aside and the keywords are highlighted instead.

Consequently, this study provides an account of the general keywords of the corpus, how
they are distributed throughout the different areas of knowledge composing the domain of
telecommunications, includes the key-keywords, that is, “the words that are key in a large
number of texts of a given type” (Scott, 1997), and narrows down the scope to the level of
a subject and a lecture in particular.

(103)
Ribera, Josep (Universitat de València, Spain): Demonstratives in the translating
mirror. An approach to translation of demonstratives from Catalan into English

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

Demonstratives are usually described as prototypical situational space deictics, but corpus
analysis shows that situational deixis is not the most frequent function that they perform.
This fact has been widely attested by Halliday & Hasan (1976) and Ariel (1990) with
respect to different textual typologies and various genres, including narratives of fiction.

Our previous research on the use and translation of demonstratives in narratives from
English into Catalan (cf. Cuenca & Ribera, 2011; Ribera & Cuenca, 2013) confirms that
non-situational uses are more frequent. Our analysis also led to establish four translation
strategies, namely:

a) Maintenance: the deictic reference of the source text, either proximal or distal, is kept in
the target text.

b) Shift: a proximal demonstrative is translated by a distal one or vice versa.

c) Neutralization: a demonstrative is translated by a non-deictic unit, implying a loss of the
deictic force of the source text in the translation.

d) Overmarking: a non-deictic unit is translated by a demonstrative, thus leading to the
introduction of deictic force in the target text.

We concluded that English and Catalan demonstratives in source and target narrative
texts, respectively, exhibit differences in use and frequency that do not match the
differences in their respective deictic systems. In fact, the maintenance and the shift of the
deictic center according to systematic correspondences between deictic systems are not
the most frequent strategies used to translate demonstratives, since non-situational text-
deictic demonstratives tend to be neutralized in the translation.

This paper aims at determining to what extent the situation differs when Catalan is the
source language. The analysis is based on a corpus consisting of the novels Mirall tren.cat
by Mercè Rodoreda and Cami de sirga by Jesús Moncada (Catalan SL), and their translation
into English (TL). We contrast both qualitatively and quantitatively whether the same
general translation patterns established in our previous research are followed or not.

The results show that the translation of situational demonstratives is more directly
explained by the different distribution of the deictic space between Catalan and English.
Specifically, higher frequencies of shift from Catalan proximal demonstratives to English
distal demonstratives are expected in connection to the addressor’s reference to the
addressee’s space. On the other hand, non-situational uses of demonstratives, mainly text-
deictic uses, are more frequent as well. However, considering that English is a full-subject
language, a lesser frequency of demonstrative neutralization is foreseen.

References


(104)

Rodríguez Martín, Gustavo Adolfo (Universidad de Extremadura, Spain): Catchphrases and characterization in Bernard Shaw’s plays: A corpus-based study.

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORAS

The dramatis personae of the Shavian dramatic canon lists several hundred characters. Many of these, regardless of their textual life and the number of lines they speak, are characterized through a skillful use of repeated discourse: catchphrases, idiosyncratic turns of phrase, and the like.

The purpose of this paper is to identify and analyze those n-word structures to establish how Shaw makes use of recurrent linguistic patterns to sketch impressionistic portraits of the minor characters and to add to the overall image of the best-rounded ones. This analysis will be carried out by investigating the lexical clusters attached to particular characters within individual plays.

(105)

Rodríguez-Puente, Paula (University of Cantabria, Spain): Phrasal verbs in the spoken language of the past: Formal and stylistic features

PANEL: CORPUS AND LINGUISTIC VARIATION

This paper analyses the formal and stylistic features of phrasal verbs in the Old Bailey Corpus (OBC) and compares them with the results of previous research carried out in ARCHER (A Representative Corpus of Historical English Registers). Phrasal verbs, combinations of a verb plus an adverbial particle which function as a single unit to various degrees (e.g. fade away, give up, turn out), tend to be associated with spoken colloquial registers, not only in PDE (see, e.g. Biber et al. 1999: 408, 409), but also in previous stages of the language (see, among others, Claridge 2000: 185-197, Hiltunen 1994, Kytö & Smitterberg 2006, Smitterberg 2008). This statement has been lately challenged by Thim (2006, 2012), who argues that in Early Modern English the (non)-occurrence of phrasal verbs in a particular text seems rather motivated by its contents, which may prompt the
use of phrasal verbs to convey literal meanings predominantly, whereas the degree of formality is a secondary aspect.

Researching the spoken language of the past has been made possible thanks to the creation of computerised corpora which contain text types that show various degrees of approximation to speech. One of such corpora is the OBC, which includes trial proceedings from 1720-1913. The Proceedings of the Old Bailey contain verbatim passages which “are arguably as near as we can get to the spoken word of the period,” (Huber 2007) thus offering the opportunity to analyse the everyday language of the past. The aim of the presentation is twofold. On the one hand, I intend analyse the stylistic and formal features of phrasal verbs in the spoken language of the LModE period through the Proceedings of the Old Bailey. On the other hand, I will draw a comparison between the use and features of these constructions in trial proceedings and other text types which also show a certain degree of ‘speechlikeliness’, such as diaries, journals, personal letters and dramatic plays.

References:


(106)

Roytberg, Anna (HSE NRU; IMB RAS, Russian Federation) & Khachko, Denis (IMB RAS, Russian Federation): Bridging corpus for Russian

PANEL: CORPUS-BASED GRAMMATICAL STUDIES
We consider bridging as special anaphoric relations where linked nominal phrases aren’t coreferent. [Clark, 1977]

Automatic Bridging resolution systems and corresponding corpora usually are devoted to some special classes of bridging relations rather than, but not to all bridging cases. Typically considered bridging relations are defined in semantic terms e.g. mereological relations.

This, in turn, undermines the existence of thesaurus, ontologies, etc. to enable recognition of the type of semantic relations in use.

Automatic bridging resolution systems usually rely on ontologies, thesaurus etc. e.g. WordNet [Vieira, R. & M. Poesio 2000; Hou et al., 2013]

Unfortunately, no sufficiently complete open source ontologies or thesaurus exist for the Russian language. This forced us to use a syntactic, rather than (not semantic) restrictions to define the considered class of bridging considered.

Our corpus is devoted to bridging in genitive construction, where; namely that is the construction where an anchor of bridging may haves the dependent NP in genitive case, but is lacking this NPhasn’t it because the potential dependent NP was used earlier and is still actualised in the mind of the reader.

An example of such construction is:

(1) В автобусе начался пожар. Водитель {автобуса} сам потушил огонь.

In bus started fire driver {bus} REL-PR-3d-S put out fire

‘The fire broke out in the bus. The driver put out the fire by himself’

The genitive construction is very common in Russian; our preliminary investigations have shown that bridging in genitive construction is a very common and important case of bridging in Russian. It occurs 2-3 times on average in an arbitrary text of 150-300 words in length. In arbitrary text of 150-300 words it occurs 2 or 3 times at the average.

Our corpus is organized as relational database and consists of news texts; each text contains 150 – 300 words. The length of texts was chosen to avoid difficulties with annotation of bridging in relatively long texts.

The corpus was annotated manually based on BRAT [Stenetorp et al., 2012]

There are no articles in Russian, therefore all nominal phrases were considered as potential participants of bridging relations.

The corpus contains about 500 texts, more than then 70 000 words, and about 1000 cases of bridging relations in genitive constructions. About 1/7 of the texts did not contain genitive bridging relations at all.

The corpus is available by request from the authors.

We plan to use the corpus as a training data set for the now developing system of automatic resolution of genitive bridging (currently in development). Additionally, we plan to grow, and to enlarge the corpus to make it suitable to test our resolution system.


(107)
Romero-Barranco, Jesús & Calle-Martín, Javier (Universidad de Málaga, Spain): Synchronic and diachronic variation: the early modern component of the Malaga Corpus of Scientific Prose

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

The Malaga Corpus of Late Middle English Scientific Prose is a research project developed at the University of Málaga in collaboration with the universities of Glasgow, Murcia and Jaén, with a twofold objective: a) the preparation of digital editions of hitherto unedited scientific Fachprosa written in the vernacular in the period 1350-1500; and b) the compilation of an annotated corpus from this material displaying the lemma, word class, accidence and meaning. This project is now in its final stage of development with a total of 1,500,000 words, the corpus allowing the researcher both word- and lemma-based queries (Calle-Martín and Miranda-García 2011: 3-20).

The implementation of the early modern English component of the corpus is then taken to be a must in order to make the corpus valid both for synchronic and diachronic variation. The present paper presents the Málaga Corpus of Early Modern English Scientific Prose with the following two objectives. The first part discusses the manuscript selection criteria together with the principles adopted for the digital editions. The second, in turn, describes the rationale adopted for the building of the corpus. Some sample searches, both word- and lemma-based, n-grams also included, will be carried out to show the potential of this corpus for linguistic research.

References:

(108)
Rosca, Andrea & Baker de Altamirano, Yvonne (Centro Universitario de la Defensa, Spain): Tracking down phrasal verbs: the case of UP and DOWN

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

This study examines the frequency and use of phrasal verbs with ‘up’ and ‘down’ in a criminal context. For such purposes, we have compiled a corpus of spoken dialogues taken from the script of the American TV series Castle shown on ABC since 2009. The corpus has
been stripped of stage directions, character names, and all incidental language, leaving a total of 210,319 words of running text.

Building on the work of McCarthy and O’Dell (2004) on phrasal verbs related to crime, in our own research we decided to enlarge the scope from purely criminal actions and look at how they are used by the police in their investigative work. We divided the phrasal verbs into two categories, namely those related or unrelated to the context of crime, which left us with a total of 187 and 409 instances for ‘up’ and 93 and 51 for ‘down’ respectively.

In the case of ‘up’, we focused on the top five phrasal verbs, which are: pick up, end up, turn up, clean up and cover up, whereas for ‘down’ they are run down, track down, narrow down, take down and go down.

Following Rudzka-Ostyn (2003), we found that the phrasal verb pick up in our corpus has three basic meanings: reaching a goal i.e. arresting/interrogating someone (e.g. Maybe we ought to pick Chloe up to see if she backs the story); (higher) up equals greater visibility i.e. capturing on film (e.g. Cam never picks her up again); and reaching the highest abstract limit or boundary, i.e. destroying the evidence after a crime (e.g. It takes real presence of mind to put five bullets into a man’s chest and then keep your cool long enough to pick up after yourself). Regarding the verb run down, only one meaning was identified, namely moving from a higher to a lower position, i.e. quickly going through a list in a database (e.g. Let’s run down reports of all stolen vehicles from the past 24 hours). A special mention should be made of the fact that some phrasal verbs are closely related in meaning and almost interchangeable in certain contexts. In the case of crime, two sets of verbs emerged in our corpus that would seem to fall into this category viz. kick up and stir up as in ‘create trouble’, and dig up and dredge up meaning ‘unearth evidence’. Finally, contrary to Krzeszowski’s hypothesis (1990) that goals are always positively loaded, we discovered that ‘up’ can sometimes have negative connotations, as in ‘end up dead’, while ‘down’ can be associated with positive outcomes, e.g. ‘track down (information)’.

Bibliography


(109)

Ruano, Pablo (Universidad de Extremadura, Spain): Charles Dickens’s Hard Times in Spanish: A Corpus-Based Approach to Speech Verbs in Four Different Translations.

PANEL: CORPORAS, CONTRASTIVE STUDIES AND TRANSLATION

By bridging the gap between corpus linguistics and literary stylistics, corpus stylistics has made new avenues of analysis available for the study of literary authors (cf. Mahlberg 2007). Despite some scholars’s initial reluctance to accept these statistical approaches, the truth is that they satisfy that demand for empirical evidence that, when it comes to the demonstration of certain aspects, can only be at best hinted at by an attentive reading. One of its major contributions is “to further our understanding of the linguistic units in literary
texts and the effects these have on the way in which readers create meanings from texts” (Mahlberg et al. 2013: 36), since such approaches may reveal new "patterns that we as readers may not be aware of” (Mahlberg 2013: 27). With the passing of time, these analyses have proved useful within the field of translation studies, for, as has been demonstrated by scholars such as Baker (2000), Xiao (2010) or Laviosa (2002, 2011), corpus methodologies can be an effective device when it comes to assessing systematically how certain elements are rendered from one language into another. Using a corpus-based approach, this study will explore an important aspect in Dickens's novels in terms of characterization: speech verbs. More precisely, an analysis of this element will be conducted in his tenth novel (Hard Times) and its translation into Spanish in four different versions. As will be seen, there are certain verbs attached to either male or female characters exclusively, and even to single characters, which results in a subtle device in terms of characterization. Nevertheless, such accuracy is not always preserved in Spanish, which may affect the way readers perceive characters through their way of speaking. By systematically comparing how this element is dealt with in four different texts, some light will be shed on the importance of keeping this precision intact so that no nuances are lost in the translated version.

Bibliography


(110)

Ruffolo, Ida (University of Calabria, Italy): The greening of hotels in the UK and Italy: A cross-cultural study of the promotion of environmental sustainability of two comparable corpora of hotel websites

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

The concern for the preservation of the environment which has been dominating all public spheres since the 1980s has led to a progressive greening of the consumers (Howlett & Raglan, 1992; Banerjee, Gulas & Iyer, 1995; Hansen, 2002). Indeed, since environmentalism has become a core value in our society, businesses and industries have been faced with public pressure to become more proactive in the protection of the environment without, however, losing profits (Harrè, Brockmeier & Mühlhäusler, 1999;
Mühlhäusler, 2003). This has led to new types of communication and different discourses through which organizations promote values and actions that aim at protecting the natural environment and achieving sustainability. However, these discourses vary across cultures since attitudes and values are transmitted by linguistic choices. As Spinzi claims (2010:19) "cultural orientations influence the way people perceive, relate to, and construct their 'environment' and 'nature' in the discourse of ecotourism", whereas, languages rely on "different linguistic choices and communicative styles to convey that particular ideological positioning".

The tourism industry has certainly not been immune to the demand of environmental responsibility, thus having to put into practice actions that do not harm the environment and employing a discourse that highlights the greening of the corporate consciousness. Various studies have been conducted to understand how tourism companies and organizations provide information on sustainability (Burman & Parker, 1993; Pritchard & Jaworski, 2005, Gössling & Peeters, 2007; Spinzi, 2010).

In light of these remarks, this paper aims at analyzing the discourse used by hotels when promoting the green practices put forth by the company. In order to analyze the effectiveness of the discourse of tourism advertising, it is necessary to investigate the link between language, text and social relations, taking into account the context of production and reception, that is all the actors and features involved in the communication: who produced it, why, who is responding to it, what social and cultural factors may influence these texts (Ruffolo, 2015).

Thus, a comparable corpus of hotel websites in British English and Italian were investigated from a cross-cultural perspective for translation purposes, focusing on the strategies adopted (and that should be adopted) by British and Italian tourist accommodation facilities on their websites.

The methodological approach of the study involves an integration of Corpus Linguistics and Discourse Analysis which provided both quantitative and qualitative perspectives. To this end, the different communication patterns and lexical choices employed within the two sub corpora were investigated in order to identify and uncover linguistic patterns and the related social and cultural features.

Thus, this paper will illustrate the features of the two subcorpora, with a focus on the significance of the node words chosen for the analysis in both languages on the basis of frequency criteria and of collocational profiles.

References


Tourism, 15/4, 402-416.


(111)

Ruiz Tinoco, Antonio (Sophia University, Japan): ueísmo y dequeísmo en Twitter, uso y distribución geográfica

PANEL: CORPUS AND LINGUISTIC VARIATION

Los fenómenos lingüísticos del español llamados queísmo y dequeísmo, consistentes en el uso indebido de una preposición, generalmente de, con algunos tipos de verbos, han sido investigados frecuentemente desde diversos puntos de vista, especialmente en sociolingüística y dialectología. Aunque de manera aproximada, también se ha señalado su extensión geográfica, principalmente en el español atlántico. Debido a la enorme extensión geográfica de los países hispanohablantes, hasta el momento no se ha llevado a cabo ningún estudio sincrónico que incluya a todos los países, tanto las zonas urbanas como las rurales.

El objetivo de este estudio es analizar dichos fenómenos basándose en un corpus de algo más de 10 millones de mensajes geocodificados (tuits) recogido aproximadamente desde junio a septiembre del año 2014. Para ello se ha utilizado el API ver1.1 Streaming de Twitter y han sido almacenados en una base de datos MySQL para posterior análisis. La base de datos, además del breve texto de los tuits, contiene las coordenadas del lugar desde donde se emitió el mensaje, y metadatos de los usuarios. Ya que no se ha restringido la zona geográfica de procedencia de los mensajes, el corpus contiene ejemplos de uso de toda la extensión geográfica del español, tanto peninsular como atlántico, de zonas urbanas y rurales.

Los verbos analizados incluyen alegrarse, preocuparse, pensar, opinar, creer, considerar, decir, comunicar, exponer, temer, advertir, avisar, dudar, acordarse, arrepentirse, etc.

Debido a que la detección de idioma de Twitter es aproximada, ha sido necesario eliminar una pequeña proporción de datos en otras lenguas como el portugués, italiano o catalán, así como los que han sido reenviados, de contenido publicitario, etc.

Para visualizar la distribución geográfica de estos fenómenos, se ha utilizado principalmente QGIS ver.2.6, un software de Sistema de Información Geográfica de código abierto. QGIS permite la preparación de varios tipos de atlas lingüísticos: por puntos, temáticos, mapas de calor, etc. También nos permite seleccionar los atributos de los datos que queremos visualizar, ya sea por el contenido de los textos, expresiones regulares, coordenadas, regiones, e incluso por los atributos o metadatos de los usuarios. Asimismo, mediante el uso de plug-ins, como qgis2leaf, los mapas confeccionados se pueden exportar a cualquier servidor estándar en Internet sin mayor complejidad técnica.
Para obtener el nombre de la ciudad, la región y el país de procedencia de los datos a partir de las coordenadas hemos preferido usar el método de matrices de distancia disponible en QGIS en combinación con una base de datos de ciudades de todo el mundo con una población de más 1.000 habitantes, de tal manera que por medio del algoritmo k-nearest neighbor se asocia cada punto a la ciudad o población más cercana.

Aunque los usuarios de Twitter no son estrictamente representativos de la población hispanohablante, hemos encontrado diferencias muy apreciables tanto del queísmo como del dequeísmo en todos los países, paso necesario para análisis posteriores.

Los atlas lingüísticos así preparados basándonos en una base de datos que contiene texto, lugar de procedencia, momento exacto de la emisión y algunos atributos de los autores, nos permiten observar visual y cuantitativamente la distribución y frecuencia de los fenómenos estudiados por lo que consideramos esta metodología muy apropiada para el estudio de la variación geolínguística.

(112)

Sánchez Berriel, Isabel (Universidad de la Laguna, Spain), Santana Suárez, Octavio,
Pérez Aguial, José (Universidad de las Palmas de Gran Canarias, Spain) & Gutiérrez
Rodríguez, Virginia (Universidad de la Laguna, Spain): Métodos para la detección de
outliers en la extracción automática de colocaciones

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

Tradicionalmente, el problema de la detección automática de colocaciones se ha resuelto recurriendo a la evaluación de indicadores que miden la atracción que se produce entre los elementos de la colocación: la base y el colocativo, basados en la frecuencia de aparición conjunta de ambos en algún corpus textual. Aunque este enfoque sea demasiado simplista para el lingüista, aporta la herramienta básica para resolver el problema desde la perspectiva computacional. En este trabajo se analiza el uso de los datos de frecuencias de formas canónicas para la extracción automática de colocaciones a través de los resultados obtenidos al aplicar indicadores como frecuencia relativa, información mutua, z-score, t-score, test de Dunning y log-Dice en el corpus del español del Grupo de Estructura de Datos y Lingüística Computacional (GEDLC) de la Universidad de Las Palmas de Gran Canaria. Los cálculos se realizan para los tipos de colocaciones sustantivo + verbo, sustantivo + adjetivo y verbo + adverbio. Bajo la premisa de implementar una solución que requiera el menor número posible de recursos léxicos, se conjuga el conocimiento lingüístico con un planteamiento estadístico. El uso de grandes volúmenes de datos textuales presentes en el corpus manejado (300 000 000 de palabras) conduce a comparar casos en que las medidas de asociación son evaluadas sobre palabras cuyas frecuencias de uso son muy dispares, lo que conlleva órdenes inadecuados en los rankings que se extraen. Por otra parte, se plantea la necesidad de establecer un criterio objetivo respecto a los valores de corte que delimiten la frontera entre las combinaciones de interés y las que se proponen como irrelevantes por los métodos estadísticos de análisis. Se revisa el enfoque tradicional, que discrimina entre combinaciones libres y colocaciones verificando si el uso de una determinada combinación se debe o no a la casualidad, determinándose rankings o valores de corte a través de indicadores que se apoyan en el concepto de independencia estadística. Se propone también una nueva estrategia que capta la característica de la preferencia de uso de una determinada combinación como elemento clave para distinguir entre colocaciones y combinaciones libres. Esta solución se basa en los métodos que se usan en estadística para detectar valores atípicos, outliers, en la terminología estadística. Además, estos se aplican considerando que del corpus surge una muestra para cada palabra, lo que permite evitar las distorsiones que se producen...
respecto a los rankings que consideran el corpus como una única muestra. Se pone de manifiesto en los resultados que este método impide que en los catálogos obtenidos en grandes corpus textuales se incluya una gran cantidad de combinaciones libres que provienen de la comparación de palabras con frecuencias de uso muy dispares. Se realiza el experimento también sobre un corpus con una cantidad de palabras mucho menor, exponiéndose los resultados de forma contrastada con los que se obtuvieron con el corpus completo. Las conclusiones y aportaciones que se derivan dan respuesta a la extracción de colocaciones de un corpus textual sea cual sea su volumen, además de aportar un punto de corte objetivo y fácilmente automatizable.

Bibliografía


(113)

Sánchez Calderón, Silvia (University of Valladolid): Is there any difference between ‘She gave a book to her daughter’ and ‘She gave her daughter a book’? English-Spanish bilingual children's acquisition of ditransitive constructions

PANEL: CORPORA, LANGUAGE ACQUISITION AND TEACHING

There has been a dichotomy in the literature on linguistic theory as far as the derivation of ditransitive constructions is concerned. Particularly, and regarding English structures,
Larson (1988, 1990) has claimed that double object constructions are transformationally derived from to-dative structures; hence example (2) derives from (1). Alternatively, Aoun and Li (1989) have stated that to-datives are derived constructions from double object structures; hence, example (1) derives from (2).

(1) a. She gave a book to her daughter [to-dative]
b. theme beneficiary thematic roles
c. accusative dative syntactic cases

(2) a. She gave her daughter a book [double object]
b. beneficiary theme thematic roles
c. accusative accusative syntactic cases
structural inherent

More specifically, the following divergences in the syntactic and semantic derivation of ditransitive constructions can be pointed out:

(a) Following Baker’s (1988) UTAH (Uniformity of Theta Assignment Hypothesis), which states that “the identical thematic relationships between items are represented by identical structural relationships between these items at the level of Deep-structure” (p.46), to-dative and double object constructions have a common underlying structure. Thus, examples (1) and (2) have the same thematic distribution, as indicated in (1a) and (2a).

(b) Conversely, double object constructions and to-datives present an asymmetry when Case theory is considered. Under locality conditions, the verbal head in to-dative constructions assigns Accusative Case to its internal DP; similarly, the preposition “to” assigns Dative Case to its adjacent DP. This is illustrated in (1c). In double object constructions (2c), the verbal head assigns Accusative Case to its internal adjacent DP. As for the second internal DP, it has inherent Accusative Case as an exceptional case marking (ECM) structure.

Taking these previous theoretical accounts, the following possibilities could occur in acquisition data when focusing on the order of acquisition of these structures:

1- Considering Baker’s (1988) UTAH hypothesis, it is expected that double object datives and to-datives have a concurrent acquisition since both structures imply a common theta-role underlying structure.

2- Concerning Case theory, double object structures are derived from to-datives, as the case that the direct object has is inherited from the direct object in to-datives. Consequently, double object structures as derived structures are expected to be acquired earlier.

3- However, previous works on the monolingual acquisition of these structures show that double object constructions appear earlier than to-datives (Snyder and Stromswold 1997), which suggests that the syntactic complexity of double object structures may not be so for children.

4- If input is taken into account, the order of acquisition could correlate with the frequency with which a child is exposed to double object and to-dative constructions.

In order to provide information on the relative order of acquisition of these 2 structures and in order to consider the four issues above, we focus on the English ditransitive verb “give” and analyze data from the CHILDES database (McWhinney 2000). We consider both child and child-directed data in the case of three English/Spanish simultaneous bilingual
children from the Deuchar corpus and the FerFuLice corpus.

Our results show that double object structures are acquired earlier than to-dative constructions and that input also plays a crucial role in acquisition since a correlation appears between the frequency of use in child-directed speech and the frequency of production in the children’s utterances.

REFERENCES


Sánchez Ibáñez, Miguel (Grupo NeoUSAL, Universidad de Salamanca, Spain): La configuración de corpus textuales para el análisis de la dependencia terminológica entre dos lenguas: retos y premisas.

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION


El objetivo de esta comunicación es proponer una metodología de configuración de corpus textuales para el análisis de la dependencia terminológica (García palacios y Humbley, 2012, Sánchez Ibáñez, 2013). Para llevar a cabo el estudio del trasvase de vocabulario especializado de una lengua en posición de hegemonía hacia otra con un ámbito de uso menos extendido se hace necesario reinterpretar el concepto de corpus. Un corpus configurado para el estudio de la dependencia terminológica puede llevar la etiqueta de “comparable”, al estar compuesto por textos en dos lenguas diferentes, pero difícilmente podrá llevar la de “paralelo”, ya que precisamente la asimetría existente entre los dos idiomas estudiados es el motivo y la razón de ser del conjunto del estudio. Todo el proceso de detección de textos, y del establecimiento de parámetros para la selección de la
información relevante para el análisis posterior vienen motivados directamente por un objetivo final: el estudio de la subordinación lingüística de un código con respecto a otro.

Teniendo en cuenta todo lo expuesto, y basándonos en un estudio concreto del trasvase de vocabulario especializado de inglés a español en el ámbito de la Enfermedad de Alzheimer (Sánchez Ibáñez, 2013), consideramos que la especificidad temática, la inmediatez cronológica del material compilado y la justificación de la elección del par de lenguas para el estudio son elementos fundamentales que no se pueden perder de vista al constituir un corpus de estas características.

En definitiva, el proceso de configuración del corpus es un paso decisivo en el desarrollo del análisis de la dependencia terminológica. Si se lleva a cabo teniendo en cuenta las características mencionadas, permite definir un conjunto de textos que combine la coherencia en ciertos aspectos como la tipología textual, el ámbito de especialidad en el que se desarrollan o el período de tiempo en el que se hayan redactado, con la diferencia manifiesta en otras cuestiones, como los cánones de prestigio y legitimidad a los que se adscriban o la intención comunicativa de sus productores. Una compilación que, en última instancia, puede ser un reflejo, en el plano textual, de los potenciales desequilibrios y diferencias que se tenga la intención de constatar en el plano terminológico.

Referencias


García Palacios, Joaquín, y John Humbley. 2012. «En torno a la dependencia terminológica». Hermeneus 14: 133-165.


(115)

Sánchez Nieto, M. T. & Zarandona Fernández, Juan Miguel (Universidad de Valladolid, Spain): Una primera aproximación al estudio contrastivo y traductológico de los nombres frasales ingleses en las lenguas española y alemana a través de corpus

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

Tanto a la lengua española como a la alemana llegan abundantes nombres frasales de la lengua inglesa o derivados de los llamados phrasal verbs compuestos de una base verbal más una partícula adverbia o prepositiva (p. ej. start-up, spin-off, etc.) que se integran en
The Lingála corpus: a tool for designing a language of instruction

(PANEL): CORPUS DESIGN, COMPILATION AND TYPES

Lingála is the most common language in Congolese music, theatre, films, radio and TV. For these reasons among others, Lingála has been spreading much more rapidly than its national counterparts (i.e., Kikongo, Kiswahili, and Ciluba). Lingála is now the most widespread language of daily communication both in Congo-Kinshasa and Congo-Brazzaville. However, like most African Languages, Lingála is a relatively less documented language (less than 1000 books published to date). Most texts are religious texts. There are also some schoolbooks, novels, comics and translations of some reports.

The reality of African schools calls for us to go beyond mere description of languages as used in society and to opt for a more prescriptive approach which could be conducive to pragmatic language standardisation, more efficient strategies for coining new terms and the production of elaborate texts in African languages. Thanks to its statistical data, corpus-based work can be a useful tool for this.

Our work of compiling a Lingála corpus aims to extract more vocabulary, to find syntax strategies of disambiguation, and identify productive and unproductive morphosemantic structures. These data will allow researchers to create efficient dictionaries, schoolbooks and to coin new terms. The final objective of this work is to allow Lingála to be better used as a language of instruction.

Compiling a Lingála corpus means dealing with a problem of language variation. Lingála is commonly acknowledged to have three main varieties: (1) Lingála lya Mankanza (LM) considered as classic; (2) Current or Spoken Lingála (CL) spoken in the northern regions; (3) and the variety which we are going to refer to in this paper as Lingála ya leló or LL (today’s Lingála), spoken in both Kinshasa and Brazzaville. LL, which is more user-friendly
than LM for most Lingála speakers, as shown by Sene Mongaba (2013), is often spoken in the so-called Lingála Facile (LF) register, i.e. a kind of code-mixing LL-French with over 20% of its lexicon constituted by French words (loanwords or code-switching). However, the presence of several French phrases or sentences in the natural language production makes it difficult to process the corpus.

To support our approach, we have analyzed a written corpus and an oral corpus. The written corpus is made up of three groups of texts: (1) religious books (the ecumenical Lingála Bible published in 2004 and the Watch Tower Bible in Lingála), (2) novels and various nonfiction writings, and (3) internet pdf and html documents. The oral corpus is also made up of three groups: (1) interviews, (2) audiovisual internet elements and (3) internet text from forums and social networks.

We have used Unitex software (Paumier 2003) to process the corpus composed of 160 652 sentences, 6 081 323 tokens and 112 826 types. In our communication, we describe our work of corpus design, compilation and types and a set of lexical patterns and morphological filters which we have used to extract data (lemmas, lexicons, grammatical construction, morphosemantic structures, etc.).

(117)
Skorczynska, Hanna & Carrió Pastor, María Luisa (Universitat Politécnica de València, Spain): Variation of general meaning key words in press releases from British and Spanish companies: gaining deeper insights into corporate discourse

PANEL: CORPUS AND LINGUISTIC VARIATION

This study compares the use of general meaning key words in the press releases of energy companies from Britain and Spain in the corpora compiled to this end. The analysis of general meaning key words in this type of specialist corpora allows for a more refined corpus-based comparison of corporate discourse than contrasting wordlists reflecting similar technical issues and terms. The focus on general meaning key words can also assist in identification of discourse strategies reflecting not only company strategic aims, but also a broader social context. Research on press releases has so far been concerned with the definition of this genre and its communicative functions (Catenaccio, 2008; McLaren & Gurai, 2005), the use of promotional language (Pander Maat, 2007; Vandenberghe, 2011), forward-looking statements (McLaren-Hankin, 2008; Vandenberghe, 2011), and rhetorical framing (Wickman, 2014). No study, to our knowledge, has adopted a corpus-based quantitative approach to focus on discourse strategies used in press releases to build and promote a corporate image. Two corpora of approximately 120,000 words were used in this study: one made up of press releases from British Petroleum and Centrica, and the other from Repsol and Iberdrola. The reference corpus used in the identification of key words accounted for nearly 1 million words and contained articles from business periodicals as well as business research papers. The key words were identified with WordSmith Tools (Smith, 2005), and they were further analysed with the Word sketch tool of Sketchengine to identify the grammatical and collocational patterns in which they were used in the two corpora. The general meaning key words included non-technical lexical items, such as ‘agreement’, and grammatical words, such as ‘the’ or ‘will’. The technical items included both the terms related with the energy production sector and with business management. The analysis of the top 50 key words has shown that the British corporate press releases contain slightly more general meaning words than the British: 25 as compared to 21. There were only three overlapping items: ‘project’, ‘programme’ and ‘agreement’. A further examination of grammatical and collocational patterns of the overlapping key words has revealed that there are significant differences in
which they were used in the two corpora. The general meaning key words with the highest keyness value in the British corpus was 'our' and 'we', while in the Spanish corpus, 'the' and 'euros'. Additionally, the top key words with the general meaning in the British corpus: 'our' and 'we' were also searched for in the Spanish corpus in order to identify frequency and collocational pattern variations. The results obtained indicate notable variations in the way general meaning key words were used in the corporate British and Spanish press releases from the energy sector. The findings suggest that despite belonging in the same production sector, the corporate discourse varies substantially reflecting both different communicative strategies, but also different social and cultural contexts from which they operate globally.

References


(118)

Soneira, Begoña (Universidad de Santiago de Compostela, Spain): Designing, Describing and Compiling a Corpus of English for Architecture

PAPER: CORPUS DESIGN, COMPILATION AND TYPES

This full paper presents the CADCE (Corpus of Architecture Discourse in Contemporary English), a collection of approximately 500,000 words of written language from a range of different sources designed to represent the language of architecture in contemporary English. The work on building the corpus began in January 2007 and the whole project was completed by December 2008. It was built for the purpose of establishing a representative corpus of architecture that reflects the lexis of this particular field. This corpus tried in this respect to fill an important gap since, to my knowledge, there is not a computer tool of this nature that can be used for research and teaching purposes.

The CADCE is monolingual and is not annotated; It includes representative North-American, British, Irish, Canadian and Australian publications (North American is by far the most represented of all variants). It is also a synchronic corpus since it gathers recent texts published from 2007-2008 (an important feature for disciplinary texts according to Orna-Montesinos (2012:129). It is a specific-purpose (following Pearson’s (1998:46) description of "special purpose corpora") corpus in the line of the Corpus of Professional English and is limited to a particular subject, namely architecture, a discipline that
comprises many other related subareas: construction, urbanism, landscape architecture, building materials, green architecture, interior design, etc. Samples are extracted only from written materials of online press, mostly specialist-non-academic articles on architecture. Other subgenres were included, namely architectural review (a subgenre describes built projects usually targeted to professionals or people concerned with the architecture field), post-construction assessment, jury citation, interview, exhibition report, architecture book review, editorials, etc.

The creation of the corpus started with a careful preparation where the design principles were established, namely representativeness (size, topic, sources, level of technicality), contemporariness (current, authentic, up-to-date publications) and accessibility (online, free-accessed, computerized texts). A preliminary pilot corpus was required to provide the general guidelines and the basis for the creation of a representative corpus. This initial pilot corpus was based on electronic materials resulting in a sample of 200,000 words. A qualitative analysis of the lexis contributed to generating many linguistic insights on the printouts which in turn led to the design of the final corpus, the CADCE, with a size of 500,000 words.

The selection of the texts in CADCE was based on an enquiry forwarded to the documentation departments of all architecture associations in the directory found in the CSCAE (Consejo Superior de Colegios de Arquitectos de España) regarding the most prominent Anglo Saxon online journals. These include among others the AIA Journal, Architecture Magazine, and Architectural Review and RIBA Journal. Full texts were retrieved excluding pictures and captions. All the texts are preceded by correlative numbers and show their website addresses and dates of compilation (most of them range from January to June 2007 except for the last one, which was compiled in December 2008).

This corpus was a crucial tool used for a full description of the most salient lexical features of a specialized discourse (following Curado, 2001:273), namely Architecture English, and also for the compilation of a trilingual glossary of architecture vocabulary in English-Spanish-Galician. It may be also a relevant instrument for further researches regarding other linguistic aspects of the architecture discourse in English.

References


(119)

**Strunk, Oliver (University of Barcelona, Spain):** Lexical Frequency Profiles of Spanish and Catalan Learners of German: Adapted LFPs of learners and native speakers in spoken argumentative texts based on single and multiword units

**PANEL: CORPORÀ, LANGUAGE ACQUISITION AND TEACHING**

Vocabulary knowledge is closely related to the communicative performance in a foreign language and specifically the four language skills (Milton 2013). But even if interest in
vocabulary acquisition research has grown in the last 20 to 30 years after a large structural oriented research era, findings still have to find their way to filter mainstream language pedagogy (Schmitt 2008). The starting point for pedagogical innovation should be the description and assessment of the level of active and passive vocabulary knowledge of a learner in relation to native speakers and the description of the variables involved. In German as a Foreign Language studies (GFL), some descriptive, corpus related initiatives have been undertaken in this field in the last years, including the publication of frequency dictionaries for learners. But specialized research about vocabulary is still rare, especially regarding width and depth, the two main aspects of active vocabulary knowledge. In this paper oral text products from learners of GFL and native speakers taken from the Varkom Corpus (Fernández-Villanueva/Strunk 2009) will be compared to analyze to which extent their lexical frequency profiles (Laufer and Nation 1995) differ when using an unedited frequency list for German (Institut für Deutsche Sprache 2009). A previous study (Strunk 2014) has shown that a lemmatized frequency list (Jones and Tschirner 2006) contributes to the description of proficiency levels when equally lemmatized texts are used, but practical restrictions imposed by the lemmatization process would make this method quite unfeasible in a pedagogical context. Furthermore, specific lexical aspects of German, the need of extension of the LFP (for example, Goodfellow 2002) and the strong relationship between different lexical scores (Crossley et al. 2014) led to the inclusion of an additional level of analysis in this study: multiword units were processed along single word units, creating a double LFP that allows to confirm the validity of the results.

(120)

Stuart, Keith (Universidad Politécnica de Valencia, Spain): Problems and Possibilities of Corpus Linguistics and Sentiment Analysis in the Health Services

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

Automating the analysis of patient feedback on health services may ultimately lead to improvements in the service. Patient narratives of experiences of the National Health Service (NHS) provide rich data of the treatment received. Their experiences are written in evaluative discourse (expression of affect in text) with a persuasive function in the hope that changes are made to the service. Sentiment Analysis research focuses on opinion mining in evaluative discourse. By using Sentiment Analysis, the health services could ensure that evaluative comments were not missed. This could make a difference to the running of hospitals and the well-being of patients.

Sentiment Analysis tries to measure subjectivity and opinion in text, usually by capturing speaker/writer evaluations (positive, negative or neutral) and the strength of these evaluations (the degree to which the word, phrase, sentence, or document in question is positive or negative). The task of automatically classifying the polarity (whether the expressed opinion is positive or negative) of texts (technically, large amounts of unstructured data) at the document, sentence, or feature/aspect level can be a challenging task. In particular, there are problems with sentences such as the following:

a) Admission was haphazard although the staff were very nice but very busy. (negative–positive–implicitly negative)

b) I would have liked more information about what I can or shouldn’t do once home for the first few days, and information regarding my follow up appointment is rather vague with no number to ring if I need assurance as I live alone. (implicitly negative)

It is these kinds of linguistic subtleties that make automatic classification difficult and result in low accuracy rates (60%-70%) of automated systems when sentences [a] express
both negative and positive opinions and express implicit negativity. As Feldman (2013: 88) states, ‘there is a need for better modeling of compositional sentiment. At the sentence level, this means more accurate calculation of the overall sentence sentiment of the sentiment-bearing words, the sentiment shifters, and the sentence structure.’

This paper describes a project where a corpus of patient narratives was created in the process of the design and development of software to automatically analyze the aforementioned corpus of patient narratives. Although there is an increasing amount of research in sentiment analysis in the clinical domain (Verhoef et al., 2014), the sentiment analysis of narratives of patient experience is a relatively understudied area (Xia et al., 2009; Greaves et al., 2013).

Problems and possibilities that arose out of this project are offered as solutions to future work on linguistic data from the clinical domain with the objective of implementing improvements in the service.

References


(121)

Tang, Xiaoyan & Cao, Jing (Zhongnan University of Economics and Law, China): Automatic Genre Classification via N-grams of Part-of-Speech Tags

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS

The recurring sequences of words have long been considered as a signifier of different genres and registers by corpus linguists (e.g. Biber & Barbieri, 2007; Biber et al., 2004; Chen & Baker, 2010; Cortes, 2004), since Biber et al. (1999) observed that the internal linguistic features of lexical n-grams are different in conversation and academic prose. His idea was furthered by Gries (2010a, 2010b, 2011), which explored the n-gram frequencies among various registers with advanced quantitative methods. The previous research mainly focused on lexical n-grams. Nevertheless, n-grams of other linguistic features, such as part-of-speech, have been much less studied (except Santini, 2004). The current study is expected to examine whether n-grams of part-of-speech tags (POS n-grams) extracted from a large corpus can be a discriminator of different genres. To be more specific, while Santini (2004) only used trigrams, the current research investigates n-grams (n=1,2,3,4,5) in order to figure out which length can best distinguish genres. BNC Baby, a genre balanced sub-corpus of BNC, is employed as the resource of POS n-grams and genres. The BNC Baby consists of four genres (i.e. academic, fiction, newspaper and conversation) and is tagged with both CLAWS5 (C5) tagset and simplified POS-tags (s-POS). Since Zipf’s Law is “true for the frequency of occurrence of n-grams” (Cavnar & Trenkle, 1994), each text is
used to generate the same quantity of n-grams. The Naïve Bayes Classifier and Multinomial Naïve Bayes Classifier in Weka (Hall et al. 2009) are used for automatic genre classification, and the performance is evaluated by 10-fold cross validation. The results show that all the weighted average F-measure obtained from this study range from 0.888 to 0.962, indicating pretty strong correlation between the occurrences or frequencies of POS n-grams and genres. In general, the findings also echo the previous studies in that the bigger the n is, the better results will be achieved. It can also be observed that when the n-grams are obtained from C5 tagging, the Naïve Bayes Classifier always performs better than Multinomial Naïve Bayes Classifier. However, if the n-grams are extracted from s-POS tagging, the Multinomial Naïve Bayes Classifier tends to perform better in five out of the twenty experiments, which may results from the balance between the utilities of the features of the tag set and the information about the n-grams, which is derived from the tag set. In addition, interesting results merged when we take a close look at the F-measure of the individual genres. Of all the twenty experiments, twelve have the best prediction for conversation, seven for fiction, one for academic, and none for newspaper. Therefore, the findings invite further research with larger corpus data and a wider range of genres as well.

References:


El objetivo de este trabajo es analizar un corpus paralelo (textos en inglés y en español)

Tarp, Sven (University of Aarhus, Denmark): El análisis de corpus es una ceremonia superflua y una forma de malgastar tu tiempo y el dinero del gobierno

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

El título de esta ponencia no representa ni mi opinión personal ni el de los colaboradores del Centro de Lexicografía de Aarhus, dedicado principal pero no únicamente al mundo de los diccionarios especializados, sino que es una combinación de dos citas de Lees (1962) y Itkonen (1976), respectivamente. La lucha de ideas entre estas posiciones y las de los defensores de los corpus fue un tema importante el la segunda parte de los años 70 del siglo pasado. Las opiniones de Lees y Itkonen fueron opuestas por Bergenholtz & Schaeder (1979), entre otros pioneros de la lingüística de corpus.

Hoy en día queda evidente que los corpus son de gran utilidad en las investigaciones relacionadas con la lingüística, y también, con la lexicografía. No obstante, la ponencia defenderá la idea de que hay un grano de verdad en las muy categóricas opiniones de Lees y Itkonen, por lo menos en lo que se refiere a la lexicografía especializada, pues dentro de esta disciplina ya no se trata de luchar con los molinos de viento sino determinar las limitaciones del uso de corpus para la confección de diccionarios especializados.

A continuación, la ponencia discutirá el conocido método terminológico de “pesca” de términos y definiciones en varios tipos de corpus mostrando con ejemplos –y al contrario de los postulados de Kilgarriff (2012)– como este método muy a menudo conduce a soluciones lexicográficas insuficientes, incorrectas y hasta peligrosas, por no hablar del gasto de tiempo y dinero. Sobre esta base, y sin negar el uso de corpus para otras tareas, la ponencia propondrá otros métodos para la selección de términos (lemas) y definiciones en diccionarios especializados, cf. Fuertes-Olivera y Tarp (2014).


Tejedor Martínez, Cristina (Universidad de Alcalá, Spain) & Martín-Pérez González, Laura (DAIL Software, Spain): Creación de tesauros: herramienta TesaurVai

PANEL: CORPORA, CONTRASTIVE STUDIES AND TRANSLATION

El objetivo de este trabajo es analizar un corpus paralelo (textos en inglés y en español)
con la nueva herramienta TesaurVai(1) para crear un tesauro de los términos seleccionados. Un tesauro es un conjunto estructurado de términos que facilita la descripción sobre un ámbito temático y el tratamiento de la información del mismo. La herramienta TesaurVai(2) integra en un solo producto un potente extractor terminológico de términos de hasta siete palabras (herramienta SimpleExtractor) con un gestor de tesauros con capacidad de soportar tesauros multilingües acorde a las normas ISO 2788, ISO 5964:1985 y su versión revisada ISO 25964-1:2011. TesaurVai es una aplicación informática desarrollada con el fin de crear y gestionar tesauros que nos permite también extraer términos. Una de sus principales características es su sencillez de manejo y sus interfaces intuitivos. La herramienta (3)permite configurar aspectos de la extracción, exportar a ficheros y proceder a la selección de aquellos elementos con los que queremos trabajar (tanto manualmente como con la ayuda del módulo de extracción de términos), para crear un tesauro organizando los términos seleccionados estableciendo relaciones y categorías entre los mismos; por ello, las conexiones semánticas que ofrece pueden ayudar a dirigir al traductor al término adecuado, así como a adarar el significado. Los glosarios terminológicos que presentan los términos organizados jerárquicamente pueden ser de gran utilidad para el proceso de traducción, de ahí la necesidad de desarrollar tesauros bilingües y la importancia de esta herramienta que ayuda en esta labor.

(1)El diseño de esta herramienta ha sido llevado a cabo por el Grupo de investigación en Validación y Aplicaciones Industriales de la Universidad Politécnica de Madrid. La empresa DAIL Software SL ha desarrollado el prototipo: http://www.dail-software.com/es/


(3)Otros programas disponibles para editar tesauros son TemaTresgratuito, DomainReuser, PoolParty Thesaurus Manager, TermTreeno, Webchoir, etc.

(124)

Tsutahara, Ryo (Tokyo University of Foreign Studies, Japan): Los derivados en el sufijo “-nte”: usos activos y no activos

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

El sufijo “-nte” se añade típicamente a verbos y forma tanto sustantivos como adjetivos. Semánticamente, ambos tipos de derivados son “activos” (cf. Laca (1993) y Rainer (1999)) que se parafrasea por una oración de relativo activa “que V”. Por ejemplo, en diccionarios, la palabra “hablante” tanto en su función como sustantivo, como adjetivo, suele definirse como “que habla”.

En este trabajo examinaremos si los sustantivos finalizados en “-nte” y los sintagmas nominales con un adjetivo como modificador del núcleo (como “una persona hablante”) son semánticamente equivalentes. Por ejemplo, en el caso de “un hablante” y “una persona hablante”, los dos parecen semánticamente equivalentes hasta cierto grado, pero, ¿será que siempre son parafraseables? En este artículo, para esta cuestión, realizaremos un análisis léxico-semántico de los derivados en “-nte” neológicos y mostraremos que hay un uso, o valor exclusivo, para los adjetivales. Intentaremos eliminar la influencia de la lexicalización cuanto sea posible por limitar los datos a los neologismos.
Consideraremos los derivados que se empiezan a usarse a partir del siglo XX como neológicos. Los datos utilizados se han extraído de los diccionarios y bases de datos "Diccionarios de Neologismos Online", "Neologismos del español" y "Corpus del español". Los usos de neologismos seleccionados fueron coleccionados de los corpus "Corpus del español" y "CORPES XXI".

Como se ha señalado en varios estudios, este sufijo es más productivo como adjetivizador. Más concretamente, en los bases de datos presentadas, hemos reunido 142 derivados neológicos en "-nte". De ellos, 85 son puramente adjetivales (no se observa su uso como sustantivo), 48 son híbridos y los 9 restantes son puramente nominales. Esta diferencia de productividad muestra que el sintagma de "sustantivo + adjetivo-nte" posiblemente tenga valores semánticos que no pueden tener sustantivos terminados en "-nte", una posibilidad que confirmaremos en el trabajo.

Los sustantivos terminados en "-nte" denotan varios tipos de entidades. Las entidades que se denotan tienen roles semánticos como AGENTE, CAUSA, etc. Sin embargo, aunque las entidades expresadas pueden tener varios roles semánticos, básicamente se corresponden con un sujeto de verbos de base. Por eso, los sustantivos en "-nte" son activos y se parafrasean como "que V".

Por otro lado, los adjetivos en "-nte" también modifican sustantivos que denotan entidades que corresponden a sujetos, como "crema alisante" ("crema que alisa"). En esos casos los adjetivos son activos pero se observa que también algunos de ellos modifican entidades no correspondientes a sujetos como "efecto alisante", que no se parafrasea como "efecto que alisa" sino como "efecto de alisar". En otros términos, actualmente, los adjetivos en "-nte" no se usan exclusivamente como activos sino también bastante frecuentemente como relacionales. Tras un análisis de nuestros datos, hemos confirmado que 76 de los adjetivos neológicos en "-nte" tienen este uso.

En conclusión, defendemos que el uso de los adjetivos en "-nte" no es necesariamente activo, y que el uso no activo es además bastante frecuente. De hecho, algunos derivados adjetivales como "dolarizante" se usan más como no activos que como adjetivos activos. Consideramos que esta dualidad entre activo y no activo es una de las diferencias semánticas entre los derivados nominales y adjetivales terminados en "-nte", y esto puede ser uno de los factores que diferencien la productividad de los dos tipos de derivados.

Bibliografía


(125)

**Turner, Chris (Coventry University, Spain):** Towards a New Approach to Some and Any Based on Large-Scale Corpus Analysis

**PANEL: CORPUS-BASED GRAMMATICAL STUDIES**

The standard grammar book description of some and any, which limits some to affirmative sentences and to questions which expect an affirmative answer, and treats any as the default form for other questions and all non-assertive contexts, has been the subject of criticism for decades, e.g. Lakoff (1969) and Lewis (1986). However, to date, all detailed alternative accounts of this area have been based either on native speaker intuition (e.g. Gethin 2011) or on small-scale corpus studies that focus on a limited range of uses (e.g. Aloni et al 2012). This conference paper describes part of a large-scale, on-going corpus study into all uses of some and any, which will provide the basis for a new description of this area. The main data source for the study is the Oxford English Corpus, a largely web-based corpus of over 2 billion words that covers a wide variety of subject areas, text types and dialects.

The paper focuses on three aspects of the some/any distinction: the use of some in negative sentences and other non-assertive contexts, the pragmatic effects of choosing some or any in certain question types and the use of both some and any in wh-questions, an area that is not covered in either current grammar books or previous corpus research.

The first part of my paper will provide reasons for my choice of corpus and explain the different research methods used in the study: simple queries, Corpus Query Language, analysis of wider context and statistical measures relating to frequency and collocational strength. It will briefly discuss issues that have arisen relating to search language and data interpretation and propose ways of overcoming these problems.

The second and main part will focus on the results of the study. I will use concordance lines, co-text and frequency data to present the following findings:

- Contrary to what grammar books claim, the use of some in negative sentences is not restricted to cases where it lies outside the scope of negation or means "some but not others".

- While some is normally used in negative sentences to refer to a partial or limited quantity, it is also used to perform "special" functions such as exhortation and rhetorical denial.

- Certain implicitly negative words accept some far more readily than others; the readiness with which these words can collocate with some is related to their meaning.

- The choice between some and any in conditional sentences is often pragmatically motivated.

- While some is the most common form for requests, offers and suggestions, pragmatic issues such as preserving face sometimes render any more appropriate.

- Because they can express a positive orientation, questions with some are frequently used in advertising texts and other types of persuasive writing.

- The meanings of some and any determine their distribution across different types of wh-questions: genuine information questions, rhetorical comment questions and negative proposition questions.

I will conclude my presentation by examining the pedagogical implications of my findings. I will discuss which of these findings are of most relevance to pedagogical grammars, learner dictionaries and course books and how the corpus data can help to present and
practise some and any in the classroom.

References
Lakoff, R (1969) Some reasons why there can't be any some-any rule
Gethin, A (2011) The truth about some and any and some thoughts it prompted on meanings, grammatical categories and academic grammars

(126)
Ueda, Hiroto (University of Tokyo, Japan) & Moreno-Sandoval, Antonio (Universidad Autónoma de Madrid, Spain): Letras and Números: web-based tools for research in Linguistics and Humanities

PANEL: CORPUS-BASED COMPUTATIONAL LINGUISTICS
This paper presents the new web-based versions of the applications LETRAS and NUMEROS. The former, LETRAS-Web allows the user to perform concordance operations in pre-loaded and new corpora. The latter, NUMEROS-Web, can perform statistical calculations. The online version has been developed by Hiroto Ueda from Tokyo University in collaboration with the Computational Linguistics Laboratory of the Autonomous University of Madrid. The objective of this project is to offer a free and online tool for the user to study data from corpora.

Both applications offer a full set of functionalities. LETRAS-Web is aimed for the linguistic part of the study of corpora:
1. Over 10 different available corpora, including CORLEC (spontaneous spoken Spanish), C-ORAL-JAPON (spontaneous spoken Japanese), ANDES (dialectal Spanish), CODCAR, CODEA, LEMI (diachronical Spanish), CODHER (vulgar Latin), a parallel corpus of Japanese and Spanish translations, and MAVIR (a Spanish and English collection of professional lectures transcribed). Additional corpora will be available in future versions.
2. Concordance in context and search in metadata.
3. Search based on pattern (regular expressions).
4. Frequency distribution of items.
5. The user can upload his/her own collection of texts.

The following figures show an example of the interface of LETRAS-Web. Figure 1 shows the query interface. After selecting the corpus and the desired metadata, the user can select in “Output” the desired result. Finally, the query can be introduced in “Pattern”, with a series of regular expressions to facilitate the query and make it more flexible.

On the other hand, NUMEROS-Web is the mathematical counterpart, designed for performing quantitative analyses and statistical operations on the results obtained from LETRAS-Web, or new data uploaded by the user:
1. Around 80 different operations, from basic median calculation, to more complex analyses based on matrix manipulations.
2. It is a free resource, in contrast to other available software such as The Sketch Engine.
3. It is web-based, in contrast to Wordsmith, AntConc or UAM-CT.

4. It provides access to small but curated corpora in different languages (Spanish, English, Japanese).

5. Sophisticated statistics as in proprietary software (SPSS).

The interface is similar to LETRAS-Web. The input is given in a matrix form, and in the output section the statistical operation is selected.

We believe in the usefulness of the tool, which has been developed as a user-friendly software aimed for linguists, philologists and computational linguists who do not need to have an advanced knowledge of computer science and are looking for a free and online tool for studying their corpora.

(127)
Vázquez García, Gloria (Universitat de Lleida, Spain) & Fernández-Montraveta, Ana (Universitat Autònoma de Barcelona, Spain): Las expresiones enfáticas de reciprocidad en las oraciones de los verbos simétricos

PANEL: CORPUS-BASED GRAMMATICAL STUDIES

En las oraciones recíprocas se pueden identificar en algunas ocasiones expresiones enfáticas que permiten reforzar la interpretación recíproca o desambiguar la lectura de la construcción pronominal, en el caso de que sea necesario. A veces, es una expresión anafórica como [(DET) uno/a/os/as PREP (DET) otro/a/os/as] (1a) y otras se trata de un adverbio que tiene un significado parecido, como mutuamente (1b).

(1) a. Los vecinos se robaban unos a otros hasta el vaho pútrido que respiraban (La tía Julia y el escritidor, Mario Vargas Llosa. Corpus del Español)

b. ...iban juntas a sus compras a las tiendas, a elegir colores y géneros consultándose mutuamente en materia de modas (Casa grande: escenas de la vida en Chile. Luis Orrego Luco. Corpus del Español).

Cabe diferenciar, sin embargo, entre las oraciones recíprocas de verbos simétricos o recíprocamente léxicos (como turnarse, intercambiar o luchar) y las que no incluyen este tipo de verbos (1). Así, parece que la presencia del elemento anafórico o la del adverbio son menos obligatorias en construcciones con este tipo de predicados. De hecho, autores como Bosque 1985 consideran que incluso hay incompatibilidades entre los verbos simétricos y alguna de estas expresiones. Concretamente, este autor apunta que el adverbio mutuamente, por ejemplo, no es compatible con un verbo simétrico. Cabe decir que los autores que tratan posteriormente esta cuestión (Peregrín Otero 1999, Quintana 2001 y 2013, Rodríguez Ramalle 2005), no varían los presupuestos de Bosque 1985. En otras cuestiones no hay acuerdo, ya que estos autores, y también Arellano 2004, mantienen que la expresión anafórica de (1) es aceptada por todos los verbos de esta clase, mientras que Devis Márquez 2006 defiende su incompatibilidad.

En nuestro estudio nos hemos centrado en este tipo de predicados y presentamos dos tipos de contribuciones en relación a su uso con expresiones enfáticas de reciprocidad. Concretamente, hemos estudiado el comportamiento de 90 predicados y 6 expresiones (además de las dos mencionadas, también se han estudiado estas 3: entre sí y sus variantes, recíprocamente, conjuntamente y juntos/as).

Por un lado, a falta de estudios empíricos en esta línea, se ha realizado un análisis cuantitativo usando el Corpus del Español (Marc Davies) que nos permite corroborar la poca frecuencia de dichas expresiones con los predicados simétricos en comparación con
las oraciones recíprocas de verbos no simétricos.

Por otro lado, nuestra segunda aportación consiste en demostrar que, aunque es menos habitual el uso de expresiones enfáticas de reciprocidad con predicados simétricos, todas estas expresiones son compatibles con estos verbos, contrariamente a lo que se ha venido afirmando. Aunque no todos estos verbos admiten todas las expresiones, no hay ninguna que no pueda aparecer con alguno de ellos. Así, hemos podido constatar que el adverbio mutuamente sí es compatible con buena parte de los verbos simétricos. Otro hallazgo es que, contrariamente a lo previsto, es mucho más habitual la expresión enfática entre sí y sus variantes (entre ellos...) que [(DET) uno/a/os/as PREP (DET) otro/a/os/as] con este tipo de verbos.

Como el Corpus del Español ha sido insuficiente para obtener datos, hemos realizado búsquedas a través de Google. Aunque Internet presenta algunos inconvenientes para analizar la lengua también tiene grandes ventajas: incluye gran variedad y riqueza de textos y es una forma eficiente de encontrar ejemplos. En todo caso, hemos priorizado las oraciones que pertenecen a textos de la prensa o bien a libros con ISBN.

Bibliografía

(128)
Vela Delfa, Cristina (Universidad de Valladolid, Spain) & Cantamutto, Lucia (Universidad Nacional del Sur CONICET, Argentina): Al abordaje de la comunicación digital: elaboración de un repositorio del español

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

Los corpus lingüísticos han encontrado en la Web un espacio dinámico y accesible para investigadores de disciplinas diversas que requieren datos primarios no siempre a su alcance. No obstante, hasta donde llega nuestro conocimiento, no existe todavía para el español un repositorio de corpus de comunicación digital que recopile muestras de chat, e-mail, SMS, entre otros géneros discursivos, apropiado para su estudio sociopragmático. Por ello, en este trabajo abordaremos la necesidad y la viabilidad en la constitución de un corpus o repositorio abierto de comunicaciones digitales en español a partir de la creación de una colección de datos de investigaciones particulares. Es decir, a través del proyecto CODICE se busca ofrecer una plataforma para crear un corpora de comunicaciones
digitales, abierto y colaborativo, a fin de ofrecer datos para su estudio sociolingüístico y pragmáticos. Es decir, se busca la optimización de los recursos invertidos en la recopilación de muestras de lenguas. De esta manera, se pondrán a disposición tanto datos de fuentes primarias como trabajos que aborden aspectos teórico y metodológicos sobre la comunicación digital.

Tras haber relevo la situación actual de los corpus sobre el español -y la escasa representatividad que en ellos tiene la comunicación digital- y de los corpus sobre diferentes tipos de interacciones en plataformas digitales (correo electrónico, chats, SMS), detectamos la necesidad de crear un repositorio de muestras estables de lengua que permita solventar esta carencia.

En el presente trabajo, ahondaremos en los aspectos metodológicos que conciernen al proyecto CODICE. Por un lado, durante la etapa de recolección de datos, se plantea como objetivo complementario la creación de unos estándares comunes, en lo que concierne principalmente a los factores contextuales y situacionales, a fin de facilitar los análisis sociopragmáticos. Por otro, para su disposición en un repositorio, se ahondará en los posibles caminos de preservación de los colaboradores. En tal sentido, se discutirá en torno a la pertinencia de ofrecer seguros suplementarios al repositorio.

(129)
Westall, Debra (Universitat Politècnica de València, Spain): A year of El País headlines on childhood obesity (2013)

PANEL: DISCOURSE, LITERARY ANALYSIS AND CORPORA

Over the past decade, researchers have examined how print media reports on overweight and obesity, being noteworthy the pioneering work by Lawrence (2004) for the USA and recent studies by Hilbert and Ried (2009) for Germany and Malterud and Ulriksen (2010) for Norway, among others. These studies all seem to confirm what health discourse specialists have long believed about media reporting and health news, the words of Evans et al. (2003: 215) summing it up nicely: “the body’ (our bodies) are being constructed, defined, regulated and pathologised by contemporary health discourse." To date, however, little attention has been given to the case of Spain, despite the alarming rise in Spanish school-age overweight and obesity rates (Serra-Majem et al., 2006; Sánchez-Cruz et al. 2013), the widespread concern for the future health of these children, both physical and otherwise (Puhl and Heuer, 2009) and the influence media can have on public perception and prevention (Boyce, 2007). In the case of Spain and according to the Spanish expert in nutrition, Félix Lobo “[...] los medios de comunicación en la sociedad moderna son un canal fundamental para la obtención de información y el cambio de comportamientos por los ciudadanos” (2007: 439).

This research aims to continue analyzing Spanish newspaper reporting about overweight and obesity, especially that involving children and adolescents (see Author, 2011). A specific corpus was initially complied with 182 news items, all published by the largest circulating national daily El País between 01/01/2013 and 31/12/2013. The articles and other news items were extracted from the paper’s online archives using the search expression “obesidad infantil”. This small corpus study will focus on the trends in coverage over the year as based on the headline analysis. The results should provide insight into the language used when the Spanish press reports on childhood overweight and obesity.


Author 2011. La obesidad infantil en la prensa española. Estudios sobre el mensaje periodístico 17/1, 225-239.

(130)

Zarco-Tejada, María Ángeles, Noya Gallardo, María del Carmen, Merino Ferradá, María del Carmen & Calderón López, María Isabel (University of Cadiz, Spain): Building a Corpus of 2L English for assessment: the CLEC Corpus

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

We describe the CLEC corpus, an ongoing project set up at the University of Cádiz (Spain) with the purpose of building up a large corpus of English as a second language classified according to CEFR proficiency levels. The goal of this corpus is twofold: on the one hand it will be used as a data resource for the development of automatic text classification systems and, on the other, it has been used as a means of teaching innovation approaches.

Nowadays, one of the main problems in our University, as far as granting our students with a language proficiency certificate, is concerned with the production of 2L English materials for language proficiency assessment. Students are to be provided with a proficiency level degree according to the levels described by CEFR. But, as CEFR authors say, the CEFR is deliberately atheoretical (Council of Europe, 2001) and adopts an action-oriented approach, describing language learning outcomes in terms of language use. Since then, there have been many groups, projects and research activities dealing with language testing and second language acquisition across Europe. One of the main goals has been the identification of criterial features for L2 English for each CEFR level (Salamoura & Saville, 2010), basic aim of the Cefling project (Alanen, R. et al. (2010a) or English Profile project (Hendriks, 2008 and Kurtes and Saville 2008), among others.

Thus, following Alanen et al. (2010b) and Hulstijn, Schoonen & Alderson (2010), and their
insights on SLA and language testing research, we have decided to collect data from existing language texts already classified according to CEFR levels and analyze them in terms of linguistic features (Banerjee, Franceschina & Smith, 2004; Norris, 1996; Norris & Ortega, 2009).

As Dahlmeier et al. (2013) point out, the success of statistical methods in NLP over the last two decades can largely be attributed to the availability of large annotated corpora that can be used to train statistical models for various NLP tasks. In this sense, our ultimate goal in making this corpus is to provide a linguistic resource for automatic text classification following a similar approach carried out for linguistic profiling of texts in Italian by Montemagni (2013) and Dell’Orletta et al. (2013).

Thus, our project was set up in 2011. We have developed CLEC with more than 100,000 words of grammatical English examples taken from 2L English texts already classified for the CEFR levels A1, A2, B1, B2, C1 and C2. The texts have been manually encoded and are divided in different groups corresponding to A1, A2, B1, B2, C1 and C2 CEFR levels. Our Corpus follows language-oriented criteria, not communicative criteria, since the classified CEFR texts used have been labeled according to linguistic facts.

The corpus has been annotated with additional information as metadata, so that each text has an identification, a reference to the main grammatical structures and a reference to the main language function identified in the text.

The creation of this corpus has been used for teaching innovation performance as well, since students of the English Studies grade have been involved in its construction not only with collecting material activities but also encoding sentences and annotating texts.

(131)
Wandl-Vogt, Eveline (Austrian Center for Digital Humanities, Austria), Declerck, Thiery (German Research Center for Artificial Intelligence, Germany) & Rainer, Heimo (Naturhistorisches Museum Wien, Austria): Bridiging the Gaps: Lexical Resources as part of European Infrastructures for the Digital Humanities

PANEL: SPECIAL USES OF CORPORA
This paper offers to discuss added values of interdisciplinary collaboration between lexicographers, computational linguists and botanists. It discusses the development of several research infrastructures in the framework of variational linguistics, cultural heritage and botany and presents a webservice to interconnect.

Modeling and added value in the framework of Linked Open Data is presented.

Freely accessible online results of the interconnection are presented.

Added value for the research infrastructures and the researchers is discussed.

Follow up collaboration within the framework of the European Network for electronic Lexicography (ENeL) and DARIAH-ERIC (working group on lexical resources) are presented.
Krabina, Bernhard & Wandl-Vogt, Eveline: Web 3.0 – Lexicography in cultural context: Wienerisch interactive / Viennese interactive

This paper introduces into a project proposal by the two authors, proposed to be funded in 2015 of the government of Vienna; situated at the Austrian Academy of Sciences. The project “Wienerisch interaktiv | Viennese interactive” is an add on and further development for the existing Wien Geschichte Wiki. The main goals of this project are:
1) Electronic lexicography in cultural context: developing a Semantic MediaWiki Schema for dictionaries, on the example on dictionaries relevant for Vienna, and discuss this in the European context, e.g. DARIAH ERIC, COST ENeL.
2) Semantic Web Technologies and Linked Open Data for the Humanities: Multilinguality on the example of the prototyped data with special focus on the city and languages of Vienna, e.g. due to migration or tourism.
3) Citizen Science: Introducing the Public to join in!
4) Electronic lexicography in cultural context: Protoypical enrichment of the Wien Geschichte Wiki / Vienna History Wiki with linguistic data (Viennese Regional language of the 19th-21st century). In doing so, enabling people to get aware of the cultural context of language.

In this presentation, the authors focus on the parts 1, 2 and 4 of this issues:

First, they introduce into the tool, Semantic MediaWiki, with special focus on its use for lexicography and lexicology in the Linked Data Era.
Second, they discuss a proposal for a general schema for dictionaries.
Third, they discuss ideas about Web 3.0 - Lexicography in cultural context.
Finally, a uncommon, fruitful collaboration between a scientific cooperation partner (AAS) and a non-research partner (KDZ) is presented.

References:

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

Wandl-Vogt, Eveline (Austrian Academy of Sciences, Austria), O’Connor, Alexander (Trinity College Dublin, Ireland), Theron, Roberto (Universidad de Salamanca, Spain) & Kieslinger, Barbara (Zentrum für Soziale Innovation, Austria): exploreAT! Re-thinking lexicography
PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

This paper aims to discuss the lexicography paradigm by introducing into a new project, established at the Austrian Academy of Sciences in Vienna (funding: 11.2014; project start: 4.2015).

The project "exploreAT! exploring austria’s culture through the language glass" aims to reveal unique insights into the rich texture of the German Language, especially in Austria, by providing state of the art tools for exploring the unique collection (1911-1998) of the Bavarian Dialects in the region of the Austro-Hungarian Empire. This corpus is large and rich, estimated to contain 200,000 headwords in estimated 4 Million records. The collection includes a five-volume dictionary of about 50,000 headwords, covering a period from the beginning of German language until the present (DBÖ, WBÖ).

In order to create enduring value from this resource, the project will apply open science and citizen science techniques to improve access and to leverage the crowd's wisdom. The engagement of users with the system will be the subject for mind-brain studies, and the results and records will be enriched and interlinked using the best practices of semantic content publishing of Linked Open Data on the Web of Data.

The key tasks are to:

1. Explore Austrian culture within a Pan-European and international setting, concerning both concepts of rural life of the multicultural Habsburg Empire, as well as supplementing historical and sociological inquiry with an understanding of the role and implementation of Lexicography over time.

2. Discover challenges and chances of a transformation process from a traditional lexicography project to an open cultural knowledge base and the role of lexicographic knowledge, especially with respect to automatic and semi-automatic techniques for publishing the results as five-star linked data.

3. Invite researchers, professionals, academics and amateurs, to participate, share and grow in the framework of up to date collaborative lexicography.

4. Create an open multilingual infrastructure for all to explore the world described by the corpus as documented in languages.

5. Reflect best practice for publishing multi-lingual linked open data, connecting the lexical, temporal, geographical and historical features of the corpus with the global and European knowledge web.

The interdisciplinary, international team is collaborating to reach the visaged goals by the following steps:

1. Create a web-based collaborative, multilingual infrastructure for archiving, editing, publishing and analysing non-standard data (historical, dialectal), it’s lexicographic output and it’s knowledge resources for scientific as well as amateur purposes.

2. Create links between different lexicographic data sets to foster exchange and interoperability at international level.

3. Engage the larger public in exploring and contributing to lexicographic data in a playful and educational approach.

4. Challenge further research for innovative ways to explore the data (e.g. visualisations, games).

5. Connect to other initiatives across the globe, e.g. DARIAH.EU, COST ENEl, EUROPEANA, LIDER, W3C Ontology Lexica Community Group, SCIentize, European School of Social
Innovation (ESSI),.opendataportal.at, Open Knowledge Foundation (OKFN), European childrens universities network (EUCU) and WIKIMEDIA.AT.

In this presentation we give an overall introduction into our project aims, focussing on the main issues concerning the development of collaborative, linked-open-data e-lexicography in an open science framework.

(134)

Lavid, Julia, Arús, Jorge (Universidad Complutense, Madrid) & Declerck, Bernard (University of Ghent, Belgium): Creation and multidimensional annotation of a register-diversified bilingual (English-Spanish) corpus for linguistic and computational investigations

PANEL: CORPUS DESIGN, COMPILATION AND TYPES

In spite of the increasing need for richly-annotated corpora in different languages in the Natural Language Processing community and the need for linguistically-interpreted parallel corpora in translation studies, existing corpora do not nearly reflect the complexity of linguistic knowledge we are used to dealing with in linguistic theory. Linguistic research questions are usually complex, often involving constraints and interactions between different linguistic categories or levels of linguistic description. Simple research questions can be answered on the basis of raw corpora or with the help of an automatic part-of-speech tagging (see Lavid 2008; Lavid et al. 2010), but when investigating more challenging interactions and relations, it is necessary to count on resources with multiple levels of annotation which allow the extraction of features at different levels. A remarkable example of such a resource for the German-English language pair is the CroCo corpus (Hansen Schirra et al. 2012, 2006; Culo et al. 2008) and its extension, the GECCO corpus (Kunz and Steiner 2012). Both corpora include original and parallel texts in English and German, and are annotated with multiple layers of linguistic information and aligned at word, grammatical, dausal and sentence levels. Using a similar corpus design to the CroCo corpus, the paper outlines current work on the construction of a richly-annotated and register-diversified textual database for the English-Spanish language pair, as recently started within the MULTINOT project. The presentation focuses on a number of linguistic and computational issues which are problematic and are currently being investigated within the project. This includes corpus design decisions such as the corpus structure –which includes four subcorpora: English originals (EO) and Spanish originals, English translations (Etrans) and Spanish translations (Strans)-, the number and types of registers to be included, the size of the samples and their comparability. It also discusses the types of information (annotation layers) to be encoded in the texts for a semi-automatic analysis of complex linguistic phenomena, as well as for enabling applications in areas of statistical NLP approaches. We also analyse the requirements and potentialities of the annotation tools which are currently being analysed for the manual annotation of discourse features in both languages, namely, the GATE platform (Cunningham et al. 2002) and the BRAT web annotation tool

References


