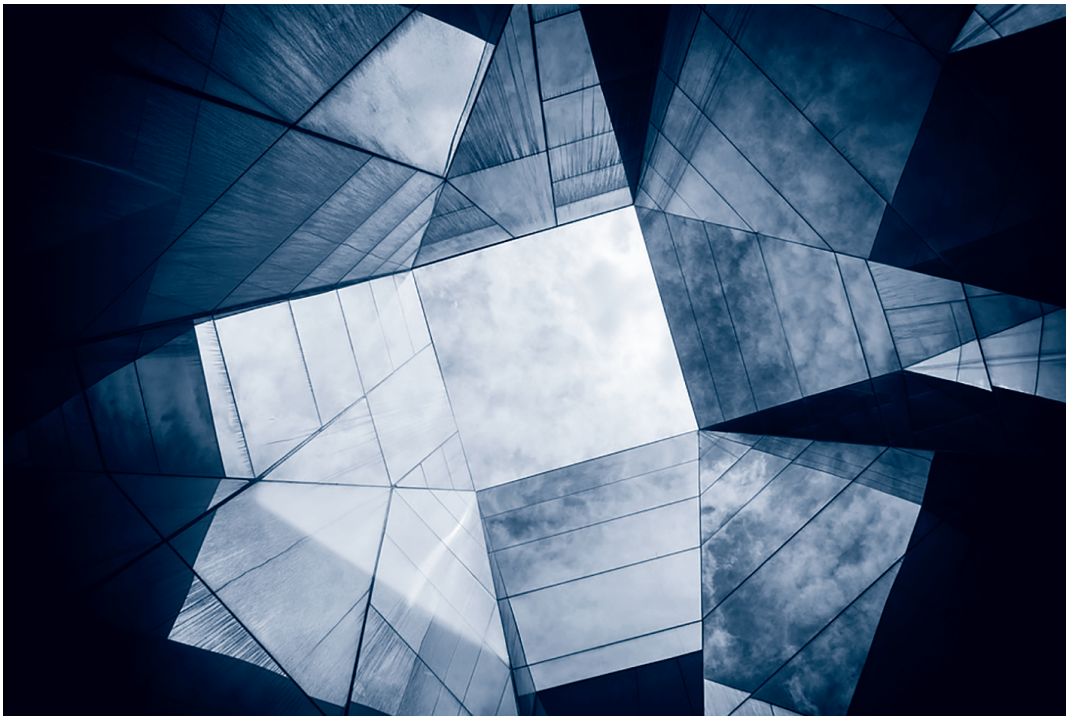
Multiperspectives in analysis and corpus design



Editores:
Miguel Fuster-Márquez
Carmen Gregori-Signes
José Santaemilia Ruiz

EDITORIAL COMARES

Interlingua

Miguel Fuster-Márquez
Carmen Gregori-Signes
José Santaemilia Ruiz
(*eds.*)

Multiperspectives in analysis and corpus design

Granada, 2020

Colección indexada en la MLA International Bibliography desde 2005

EDITORIAL COMARES

INTERLINGUA

249

ENVÍO DE PROPUESTAS DE PUBLICACIÓN:

Las propuestas de publicación han de ser remitidas (en archivo adjunto, con formato PDF) a alguna de las siguientes direcciones electrónicas: anabelen.martinez@uco.es, psgines@ugr.es

Antes de aceptar una obra para su publicación en la colección INTERLINGUA, esta habrá de ser sometida a una revisión anónima por pares. Para llevarla a cabo se contará, inicialmente, con los miembros del comité científico asesor. En casos justificados, se acudirá a otros especialistas de reconocido prestigio en la materia objeto de consideración.

Los autores conocerán el resultado de la evaluación previa en un plazo no superior a 60 días. Una vez aceptada la obra para su publicación en INTERLINGUA (o integradas las modificaciones que se hiciesen constar en el resultado de la evaluación), habrán de dirigirse a la Editorial Comares para iniciar el proceso de edición.

Polígono Juncaril
C/ Baza, parcela 208
18220 Albolote (Granada)
Tlf.: 958 465 382
https://www.comares.com • E-mail: libreriacomares@comares.com
facebook.com/Comares • twitter.com/comareseditor • instagram.com/editorialcomares

ISBN: 978-84-1369-009-4 • Depósito legal: Gr. 737/2020

Fotocomposición, impresión y encuadernación: COMARES

Table of Contents

LIST OF CONTRIBUTORS VII

INTRODUCTION XI
Miguel Fuster-Márquez
Carmen Gregori-Signes
José Santaemilia Ruiz

ADVANCE-FEE SCAMS: A CORPUS AND GENRE ANALYSIS 1
Ismael Arinas Pellón
Patrizia Anesa

EL SABOR DE LAS MANZANAS: ANÁLISIS CONTRASTIVO (ESPAÑOL-INGLÉS) DE LA TERMINOLOGÍA OBJETIVA REFERIDA A LA EXPERIENCIA SENSORIAL DEL GUSTO 15
Leonor Pérez Ruiz
María-Teresa Ortego Antón

TWO EXAMPLE MARKERS IN AND BEYOND EXEMPLIFICATION: DIALECTAL, REGISTER AND PRAGMATIC CONSIDERATIONS IN THE 21ST CENTURY 33
Paula Rodríguez-Abruñeiras

PROBABILISTIC GRAMMARS ACROSS REGISTERS: PRONOMINAL SUBJECT EXPRESSION IN SOME VARIETIES OF ENGLISH 47
Iván Tamaredo

SEMANTIC FRAMES IN SCIE-LEX 61
Isabel Verdaguer
Emilia Castaño
Natalia Judith Laso

ACCESIBILIDAD, TRADUCCIÓN AUDIOVISUAL Y NORMAS EN LA SUBTITULACIÓN ONLINE: EMPAC (EUROPARLTV MULTIMEDIA PARALLEL CORPUS) 73
Iris Serrat Roozen

FINT-ESP: A CORPUS OF FINANCIAL REPORTS IN SPANISH. 89
Antonio Moreno-Sandoval
Ana Gisbert
Helena Montoro

SPELLING NORMALISATION AND POS-TAGGING OF HISTORICAL CORPORA: THE CASE OF GUL, MS HUNTER 135 (FF. 34R–121V) 103
Jesús Romero-Barranco

ANNOTATING FACTUALITY IN THE TAGFACT CORPUS 115
Glòria Vázquez García
Ana Fernández-Montraveta

list of contributors

Patrizia Anesa is a Senior Researcher in English Language and Linguistics at the University of Bergamo and she holds a PhD in English Studies from the University of Verona. Her research interests lie mostly in the area of specialised discourse, with particular reference to the investigation of knowledge asymmetries in professional communications. She has been a member of several national and international projects in the field of ESP. Email: patrizia.anesa@unibg.it

Ismael Arinas Pellón is a lecturer at the Universidad Politécnica de Madrid (UPM). He teaches Communication for Specific Purposes to industrial engineering undergraduates. He is currently adapting materials for these courses. As a result of this syllabus adaptation, he is researching how technology is described in different contexts (patents, text-books, journal articles, marketing materials). Additionally, he is analysing persuasion strategies in different genres combining genre analysis, Maton's Legitimation Code Theory, social psychology, and phenomenology with corpus linguistics. He is a member of an Innovation on Education group known as Comm&Learn as well as of the Global Engineers Language Skills (GELS) international network. Email: ismael.arinas@upm.es

Emilia Castaño is an adjunct lecturer in the department of Modern Languages and Literatures and English Studies at the University of Barcelona. She holds a PhD in English Philology from the University of Barcelona and her research interests include corpus linguistics and the application of cognitive semantics, frame semantics, image schemas and conceptual metaphor theories, to discourse analysis. Email: e.castano@ub.edu

Ana Fernández-Montraveta holds a PhD in Linguistics and Communication and an MA in Computational Linguistics. She is currently an associate professor in the English and German Department at the Universitat Autònoma de Barcelona and leader of the Inter-University Research Group on Language Applications (GRIAL). Her research focuses on English-Spanish contrastive analysis, more specifically on the field of constructions. In the last few years, she has carried out research on aspectuality, modality and factuality, always at sentence level. Especially relevant is her participation in the creation of various resources, among which there is the Spanish WordNet 3.0, highly useful for translation between Spanish and English, and several corpora annotated with linguistic information. She co-authored the books *Clasi-*

ficación verbal. Alternancias de diátesis and *Las construcciones con 'se' en español*, and has more than fifty publications. She has led various projects related to corpus compilation and annotation, such as SenSem. She currently participates in the TAGFACT project. Email: Ana.Fernandez@uab.cat

Ana Gisbert is Associate Professor in the Department of Accounting in the Faculty of Economics at Madrid Autonomous University. She was a Predoctoral Fellow at Lancaster University within the context of the HARMONIA European project on Accounting Harmonisation and Standardisation in Europe. She has published papers in the areas of international accounting, corporate governance, audit oversight and earnings management. Her current research interests are focused on the analysis of financial reporting narratives. Email: ana.gisbert@uam.es

Natalia J. Laso is a Serra Hunter fellow in English Linguistics at the University of Barcelona. She holds a PhD in English Philology from the University of Barcelona and is also a member of the GRELIC-Lexicology and Corpus Linguistics Research Group. Her research is focused on two main areas: a) science writing and the main challenges that NNES writers face when writing their research in English; and b) the use of corpora in the linguistics classroom. She has co-edited the volume *Biomedical English: a corpus-based approach* (by Isabel Verdaguer, Natalia J. Laso & Danica Salazar. Eds.), published by John Benjamins Publishing. Email: njlaso@ub.edu

Helena Montoro Zamorano is a computational linguist, graduate in Translation and Interpreting (2017) and Master's degree in Spanish Language Research (2019) from the Universidad Autónoma de Madrid. She currently works at the Instituto de Ingeniería del Conocimiento, in the Social Business Analytics department. She participates in projects and research on Natural Language Processing, carrying out tasks -such as creation of computational grammars, elaboration of taxonomies and dictionaries, compilation and annotation of corpus or design of chatbots- related to the development of linguistic resources for the detection of sentiment in social media and the training of machine learning models. Email: helena.montoro@iic.uam.es

Antonio Moreno-Sandoval (BA 1986, MA 1988, PhD 1991, Universidad Autónoma de Madrid, UAM) is Associate Professor of Linguistics and Director of the Computational Linguistics Lab at UAM, and Director of the UAM-IIC Chair in Computational Linguistics. He is a former Fulbright postdoc scholar at the Computer Science Dept., New York University (1991-1992) and a former DAAD scholar at Augsburg Universität (1998). His training in Computational Linguistics began as a research assistant in the Eurotra Machine Translation Project (EU FP-2) and then at IBM Scientific Centre in Madrid (1989-1990). He was the principal researcher of the Spanish team in the C-ORAL-ROM Project (EU FP-5). Since 2010 he is Senior Researcher at the Instituto de Ingeniería del Conocimiento (IIC-UAM) in the Social Business Analytics group. Moreno-Sandoval has supervised 12 theses to completion. He is author or co-author of 5 books and over 80 scientific papers. Email: antonio.msandoval@uam.es

María-Teresa Ortego Antón es Profesora Ayudante Doctor en el Departamento de Lengua Española de la Universidad de Valladolid y miembro del CITTAC de la mencionada universidad, así como del grupo interuniversitario ACTRES de la Universidad de León. Colabora con el grupo LEXYTRAD de la Universidad de Málaga y con el OLST de la Universidad de Montreal. Imparte docencia en la Facultad de Traducción e Interpretación de la Universidad

de Valladolid, en asignaturas relacionadas con la interpretación, la traducción especializada y las tecnologías de la traducción en la Facultad de Traducción e Interpretación. Licenciada en Traducción e Interpretación en 2007, obtuvo el grado de Doctor en Traducción y Comunicación Intercultural en noviembre de 2012 gracias a la concesión de una Ayuda FPI. Entre sus publicaciones destacan artículos en revista, capítulos de libro y libros sobre lexicografía bilingüe, terminología, tecnologías de traducción y lingüística de corpus aplicada a los Estudios de Traducción. Ha presentado ponencias en varios congresos internacionales. Además, es miembro de varios proyectos de investigación nacionales e internacionales sobre traducción especializada e interpretación. Ha trabajado como traductora e intérprete profesional durante más de diez años. Email: mariateresa.ortego@uva.es

Leonor Pérez Ruiz es Profesora Titular de Universidad en el Departamento de Filología Inglesa de la Universidad de Valladolid y miembro del CITTAC (Centro de Investigación en Terminología bilingüe, Traducción especializada y Análisis Contrastivo). Su destino docente está en la Facultad de Filosofía y Letras donde imparte docencia en el Grado de Estudios Ingleses. Sus áreas de especialización son las lenguas con fines específicos, lenguas con fines académicos, la lingüística del corpus y la traducción. Ha sido investigadora principal en varios proyectos y también ha participado en proyectos europeos. Ha publicado numerosos artículos en revistas especializadas, capítulos de libros y monografías. También ha presentado trabajos en Congresos Nacionales e Internacionales. Email: lperezru@fyl.uva.es

Paula Rodríguez-Abruñeiras holds a European Doctorate in Historical Linguistics from the University of Santiago de Compostela (January 2015). From August 2013 to May 2015, she was an assistant of Spanish at the Department of Spanish and Portuguese of the University of Wisconsin-Milwaukee (USA), where she also completed a two-year M.A. Degree in Spanish. Between September 2015 and January 2016, Paula held a visiting lecturer position at the Department of English, French and German of the University of Vigo, and in February 2016 she joined the Department of English and German at the Universitat de València, on a position as Lecturer in English. She has also carried out research at both national (University of Vigo, University of Santiago) and international (University of Helsinki, University of Manchester, University of Freiburg, University of Wisconsin-Milwaukee) universities, and at the British Library too. She is currently a member of GENTEXT and IULMA, and an affiliated member of LVTC (http://view0.webs.uvigo.es/team/paula-rodriguez-abruneiras). Her research interests are historical linguistics, linguistic variation, new Englishes, corpus linguistics, critical discourse analysis and gender and sexual (in)equality issues. Email: paula.rodriguez@uv.es

Jesús Romero-Barranco studied English Language and Literature at the University of Málaga, where he earned two MA Degrees, one in advanced English studies (2013) and the other in teaching English as a foreign language (2014). In 2017, he received his PhD with a thesis entitled "Early Modern English Scientific Text Types: Edition and Assessment of Linguistic Complexity of the Texts in MS Hunter 135 (ff. 34r–121v)". He is a member of the Department of English and German (Universidad de Granada) and has been a visiting researcher at the Department of English Language (University of Glasgow) and the Department of the History of English (Adam Mickiewicz University, Poznań). Among his research interests, Dr. Romero is not only interested in Historical Linguistics, Palaeography and Codicology, but also in morphosyntactic variation in Present-day English and its varieties around the world. Dr Romero-Barranco has

published in journals such as Atlantis, Studia Neophilologica or English World-wide, among others. Email: jesusromero@ugr.es

Iris Serrat Roozen holds a PhD in Applied Languages, Literature and Translation by the Universitat Jaume I. She is currently a professor at the Catholic University of Valencia San Vicente Mártir. Her main lines of research include audiovisual translation, accessibility, and second language learning and teaching. Email: iris.serrat@uv.es

Iván Tamaredo holds a BA in English Language and Literature (June 2012), and an MA (June 2013) and a PhD (November 2018) in English Studies. Formerly a pre- and postdoctoral researcher at the University of Santiago de Compostela under funding from the Spanish Ministry of Economy and Competitiveness (grant no. BES-2015-071233), Iván is currently a Substitute Lecturer in English at the University of Vigo. His research interests include varieties of English, probabilistic variation, language processing, and linguistic complexity. He has presented papers at several major international conferences (*ChangE* 2015 at Helsinki, and *ICAME 36*, *37*, and *39* at Trier, Hong Kong, and Tampere, respectively) and published articles in peer-reviewed journals such as *English World-Wide*, *English Language and Linguistics*, *Atlantis*, and *ICAME Journal*. Email: ivan.tamaredo.meira@uvigo.es

Gloria Vázquez is a senior lecturer in the area of General Linguistics at the University of Lleida and a member of the Interuniversity Research Group on Language Applications (GRIAL). At the beginning of her career, she was interested in the study of verbal semantic classes and their degree of internal homogeneity in relation to syntactic behavior, which is why she studied verbal syntactic patterns together with their constructional meaning. Her contributions to the description of Spanish verbal periphrasis stand out, and in recent years, she has started a new line of research related to the degree of factuality expressed in sentences. She has more than fifty publications, among which 2 books, *Verbal classification*. *Diathesis alternations* and *Constructions with "se" in Spanish,* stand out. Other important publications are two chapters in two manuals: one on corpus linguistics and another on Spanish lexicography. In these areas, she has also participated in the creation of some large-scale resources and collaborated in several funded research projects. She is currently the main investigator of the TAGFACT project. Email: gvazquez@dal.udl.cat

Isabel Verdaguer Clavera is Professor of English at the University of Barcelona. Her research interests and publications include the history of translation, contrastive linguistics, corpus linguistics, cognitive linguistics and lexicology. She coordinates the GRELIC-Lexicology and Corpus Linguistics Research Group and has coedited *Biomedical English: a corpus-based approach* with Laso and Salazar (John Benjamins, 2013). Her recent work includes "Identifying verb collocational patterns in a specialized medical journal corpus: a pedagogical approach to phraseology" (with Noguchi) 2018, in *RESLA*, 31, 2; and "Semantic frames and semantic networks in the Health Science Corpus" in *ELiEs* (forthcoming). Email: i.verdaguer@ub.edu

Introducction

In this volume, the readers of *Multiperspectives in Analysis and corpus design* will find nine selected peer reviewed and original contributions which deal with key aspects in recent trends in corpus linguistics, such as the developments in corpus design, compilation procedures and annotation, and the different analytical perspectives in which corpus techniques have become a core empirical methodology, either in isolation, or combined with other approaches that help reinforce arguments. It will be found that, in most of the articles, the authors themselves have compiled their own study corpus. Consequently, as it is customary in Corpus Linguistics research, a justification of the compilation procedure (e.g. sampling parameters or representativeness) is part and parcel of the discussion. The research areas to which corpus linguistics has been successfully applied in this volume include historical linguistics, linguistic variation, discourse analysis, computational linguistics and translation.

The first paper by Arinas and Anesa, "Advance-Fee Scams: A corpus and genre analysis", provides a detailed analysis of online scams. The authors state that online scams have "become a global criminal phenomenon which, worryingly, is on the increase". Their corpus contains over 500 fraudulent email texts. Their study aims to identify the persuasive strategies used by scammers to manipulate their victims by generating their confidence to deceive them. A keyword analysis is applied as a crucial analytical technique followed by a qualitative analysis. The analytical frameworks applied include Bhatia'[1]s notions of promotional letter move structure, Gao & Gao's representation of knowledge involved in fraud detection and prevention, Kahnenab's model of judgment and choice according to preferences and attitudes, and Fischer, Lea and Evans' persuasion strategies. An important conclusion is that scammers "use a strong narrative within

[1] All references to authors or titles mentioned in this introduction will be located in the bibliographical references provided by each contributor to this volume.

a recognizable genre", creating the "illusion of intimacy, sincerity, and urgency", to persuade and tease their victims.

Pérez and Ortego carry out a contrastive terminological analysis around food entities, in particular, apples: "El sabor de las manzanas: Análisis contrastivo (español-inglés) de la terminología objetiva referida a la experiencia sensorial del gusto". In their paper, the authors of this second article perform a contrastive Spanish-English discourse study of terms used to name the gustatory perception. For that purpose, they make use of a comparable bilingual study corpus based on fact sheets on apples gathered from websites of Spanish and British food companies. They look into meaning choices through frequent collocates, and co-textual intensifying and mitigation strategies. They find that in their description of apple taste, these fact sheets typically appeal to sensory qualities such as sweetness or acidity, the evocation of different foods and beverages, the aroma and the touch. In their conclusion, they highlight that both languages use a wide variety of terminology to describe, with great accuracy, the different taste sensations that the consumption of an apple causes on the palate. They hope that this study can be useful for translators, who would need to use the appropriate terminology and discourse features when promoting online products.

The third article, by Rodríguez-Abruñeiras, examines the use of *for example* and *for instance* in Present-Day English in two corpora, the *British English 2006* and *American English 2006*, each of these containing one million words. The author follows Eggs and McElholm's typology of exemplification, selection and argumentation, undertaking a corpus-based study of two mainstream varieties of English, namely British and American English, in the early twenty-first century. Among other aspects, her analysis reveals that both markers show quite similar overall distributions in the two general corpora, and that the argumentative function prevails over others in both mainstream varieties of English. Regarding text types, the results indicate that fiction has very few examples of these markers, when compared with informative prose, which is due, according to Rodríguez-Abruñeiras, to their differing rhetoric and communicative purposes. Thus, she argues that "fiction is creative and is not expected to be accurate except as a reflection of human experiences, whereas text types, such as science or law (which are argumentative by nature), are characterised "by the use of more accurate and concise kind of language". Rodríguez-Abruñeiras claims that both markers "may also bring about various pragmatic nuances, such as focus or mitigation", most particularly in scientific prose. In her conclusion, she highlights that that there are no significant differences between British English and American English. In both varieties there is "clear preference for *for example* over *for instance*" and a similar use of the three functions she examined in this paper.

Tamaredo's paper "Probabilistic grammars across registers: Pronominal subject expression in some varieties of English", fourth in this monograph, takes a dialectal and variationist/sociolinguistic perspective. Tamaredo focuses on the internal and external constraints which determine the choice between overt and pronominal subjects among speakers of three varieties of World Englishes with different regional and cultural back-

grounds, namely British English, Indian English, and Singapore English. This research was carried out by exploring the corresponding three national corpora, containing one million words each, within the *International Corpus of English*. The author adopts the double approach of probabilistic grammar framework, which assumes that grammatical knowledge is at least partially experience- and a usage-based approach. The author takes into account various internal language-constraints as predictors of subject pronoun omission, and language-external factors (variety, mode of production, and level of formality). He makes use of VADIS variationist modelling, in order to examine probabilistic differences. Tamaredo's findings seem to demonstrate that, among the external factors, the mode of production and the level of formality outweigh variety to account for variation between omitted and overt subject pronouns in the three varieties. As for the internal factors, the author concludes that coordination and clause position are the most crucial constraints. As in many cases of morphosyntactic variation in PDE, Tamaredo finds that register-related factors appear "to override variety-specific patterns".

"Semantic frames in *SciE-Lex*" is the title of the fifth paper, by Verdaguer, Castaño and Laso. The paper focuses on the analysis of two English verbs, *to block* and *to inhibit,* very frequently used in biomedical discourse. The article discusses a new development of *SciE-Lex*, a lexical database of biomedical English, a development from their *Health Science Corpus*, a four-million-word corpus of biomedical English. In this new development, the authors applied Fillmore's frame semantics' approach and used the well-known database *Framenet*, which classifies individual words on the basis of their semantic frames. The aim of the *SciE-Lex* project is to help the Spanish biomedical community to publish papers in English which conform "to the conventions of scientific discourse". At present, *SciE-Lex* contains phonological, morphological, syntactic, semantic, collocational and phraseological information on highly frequent biomedical English words. The authors analyse the frame-based information for both verbs, which enables them to observe the syntactic and semantic patterns that both of them share. This approach makes it possible to integrate lexical meanings into a higher level of organization, as well as compare and contrast their preferred meanings, collocational and syntactic patterns against those in general discourse. Since the meanings of words in domain-specific texts tend to be more specific than in general use, the authors claim that their idiosyncrasies should be considered. The authors conclude that "the application of Frame Semantics to the semantic and syntactic description of the verbs *to block* and *to inhibit* in their specialised corpus bears witness to the existence of "remarkable differences between general and biomedical English".

Subtitling norms and accessibility in audiovisual translation is the focus of the sixth contribution in this monograph. Serrat's paper, "Accesibilidad, traducción audiovisual y normas en la subtitulación *online*: EMPAC (EuroparlTV) Multimedia Parallel Corpus", discusses issues of accessibility on the audiovisual content in subtitling on the online television channel EuroparlTV. Her analysis is based on EMPAC, a Spanish and English parallel (English/Spanish) corpus of 5 million words she has compiled for the years 2009-2017. In this contribution, Serrat discusses the norms which underlie the online

subtitles generated by translation professionals to determine whether they are aligned with the commitment to equal access to information for all European citizens. She homes in on the following relevant features: reading speed, pauses between subtitles, and characters per line, as well as (in)adequate segmentation of syntactic patterns. Her conclusions indicate that the lack of commonly accepted standards for online subtitling leads to malpractice and dubious quality and illegible subtitles. In her view, these are barriers for effective communication in this institutional EU context.

Moreno Sandoval, Gisbert and Montoro's "A corpus of financial reports in Spanish" provides an overview of the different steps taken to retrieve texts of annual reports in a corpus of contemporary Spanish financial narratives (FinT-esp) a corpus compiled by their research team. This is a contribution to the development of existing methodologies in computational linguistics and, more specifically, aims to help finance and accounting fields in the processing, classification and analysis of large amounts of financial narratives provided by PDF files. A great deal of the article is dedicated to the description of the contents of FinT-esp and the methodology followed for the compilation of the corpus. The authors use an adaptation and modification of the CFIE-FRSE tool developed by Mahmoud El Haj (Lancaster University), an application that detects the structure of annual reports and allows them to extract their contents at section level. Once this was done, the texts were processed with computational tools such as POS taggers or parsers. For example, they highlight that to help prospective users in their searches "[t]he texts are indexed with Elastic Search, which favours "a fast and easy search as if we were using a Google-like tool". To illustrate the usefulness of this specialised corpus, Moreno-Sandoval, Gisbert and Montoro discuss three lexical features: modality (to detect possible linguistic biases), keyword analysis (to identify domain specific vocabulary) and polarity (by applying a computational tool of sentiment analysis). Regarding possible applications, the authors state that it is already being used for terminology extraction and, claim that FinT-esp could become "a valuable source of data to facilitate both the development of new techniques in computational linguistics and the promotion of interdisciplinary studies between accounting and linguistic academics".

Romero-Barranco shows his interest in the compilation and annotation of historical corpora in his paper "Spelling normalisation and POS-tagging of historical corpora: The case of GUL, MS Hunter 135 (ff. 34r-121)". In this eighth contribution, Romero-Barranco brings in decisions or suggestions involving the process of compiling and annotating an early Modern English linguistic corpus. As a source of evidence, the author examines the criteria used to compile and annotate the Glasgow, University Library, MS Hunter 135 (ff. 34r-121v), a medical volume written in the first half of the sixteenth century. Given that early Modern English features spelling variation, it is his view that diachronicians would need to normalise the orthography if they wish to enhance the performance of an automatic POS-tagger as would be done with a contemporary text. The spelling normalisation is carried out with *VARD* (Rayson, Archer and Smith 2005). One of its most crucial features is that *VARD*, the variant detector tool, not only normalises spelling variants, but also adds a tag showing where this original

spelling is kept. Romero-Barranco also gives and overview of progress in automatic part-of-speech tagging. Most particularly he looks at what has been accomplished by developments of the CLAWS (Constituent Likelihood Automatic Word-tagging System) POS (Part of Speech) tagger. Both VARD and CLAWS taggers have been applied to this early Modern English document. The study shows that thanks to spelling normalisation, part-of-speech annotation with CLAWS increases its accuracy approximately by 15%. Particularly, CLAWS' accuracy is enhanced from 82.9% and 83.9% to 96.8% and 97.2% in the surgical text and the medical recipes. Nevertheless, Romero-Barranco points out that historical linguists who work with earlier texts need to make decisions about problematic issues such as how to deal with archaic words, genitives or compounds, among others. Additionally, he claims that automatic normalisations must be checked and manual amendments would be required.

In the last contribution to this volume, "Annotating factuality in the TAGFACT", Vázquez and Fernández-Montraveta focus on the complexity of applying automatic corpus annotation to do research on factuality and evaluation. The authors describe and give a rationale for the labels used for the annotation scheme of the corpus TAGFACT. The chapter takes into account, discusses and illustrates with numerous examples four aspects: eventual types (dynamic and non-dynamic situations) and the writer's commitment to the certainty of the assertion, polarity and time. Their analysis aims to capture "the degree of commitment of an author regarding the certainty of the facts he or she is narrating". The authors highlight three innovations which result from their approach. First, it helps distinguish dynamic from non-dynamic situations. Second, it allows the identification and annotation of absolute truths. Thirdly, it helps distinguish prototypical from eventual properties. Vázquez and Fernández-Montraveta claim that their system is the only one at present which is based exclusively on linguistic knowledge, for Spanish. Even though this model has been created for the annotation of a Spanish corpus, they claim that it may be applicable to other languages.

We must thank the Valencian Government (Generalitat Valenciana) for kindly giving financial support to this publication (project code 31/1769321).

Miguel Fuster-Márquez
Carmen Gregori-Signes
José Santaemilia Ruiz

colección:

INTERLINGUA

249

Dirigida por:

Ana Belén Martínez López y Pedro San Ginés Aguilar

The readers of *Multiperspectives in Analysis and Corpus Design* will find in this volume nine selected peer reviewed and original contributions which deal with key aspects in recent trends in corpus linguistics, such as the developments in corpus design, compilation procedures, and annotation. All the contributions in this book use corpus techniques as a core empirical methodology, either in isolation, or combined with other approaches. Furthermore, in most of the articles, the authors themselves have compiled their own study corpus. Consequently, as it is customary in Corpus Linguistics research, a justification of the compilation procedure (e.g. sampling parameters or representativeness) is part and parcel of the discussion. The research areas to which corpus linguistics has been successfully applied in this volume include historical linguistics, linguistic variation, discourse analysis, computational linguistics and translation.

We remain indebted to the Valencian Government (Generalitat Valenciana) for kindly giving financial support for this publication (project code 31/1769321).

COMARES
editorial

ISBN 978-84-1369-009-4