Setting up a Corpus-based Study of Hedging in Spanish Research Articles: Creating a Valid Corpus

Sonia Oliver del Olmo
Departament de Projectes d'Enginyeria
Universität Politècnica de Catalunya

This paper presents the steps and rationale followed in creating a Spanish corpus of biomedical research articles, comparable to the ones used to study genre patterns in English language biomedical research papers, such as that used by Salager-Meyer in 1994 to study hedging in English medical discourse. Salager-Meyer's corpus was taken as the preliminary model. It consisted of 5 research papers (RP) and 10 case reports (CR) from five high impact factor journals in English. In the Spanish corpus, twenty articles from the same two medical genres and from six different Spanish scientific journals in biomedicine were carefully selected. This paper gives a detailed description of the steps taken in the process, as well as the relevant changes in the selection criteria for the Spanish corpus.

Palabras clave: análisis de corpus, retórica contrastiva, estudios interculturales, discurso científico, lenguas con fines específicos y atenuación retórica.
Fecha de recepción del artículo: 27 de marzo de 2004

Sonia Oliver del Olmo
Departament de Projectes d'Enginyeria
Universität Politécnica de Catalunya
Secció d'Anglés. ETSEIT, TR5
Dirección: Gomera s/n 08027
Sant Cugat (BCN)
Correo electrónico: Sonia.01iver@telefonica.net

En este artículo describimos y argumentamos los pasos llevados a cabo en la creación de un corpus de artículos de investigación biomédica en español, comparable al utilizado por Salager-Meyer (1994) para el estudio del artículo de investigación biomédica en inglés. Esta autora investigó el fenómeno de la atenuación retórica en el discurso médico en inglés y nosotros tomamos su corpus de estudio como modelo preliminar para elaborar el nuestro en español. Dicho corpus consistía en cinco artículos de investigación (AI) y 10 casos clínicos (CC) en inglés, publicados en cinco revistas especializadas con un alto índice de impacto. En nuestro corpus en español, seleccionamos 20 artículos de los mismos géneros médicos que Salager-Meyer, publicados en español en seis revistas especializadas en el campo de la biomedicina. Este artículo ofrece una detallada descripción del proceso de elaboración de nuestro corpus y especifica algunos cambios sustanciales en cuanto al criterio de selección del mismo.

1. *Introduction*

Evidence for genre features of research articles comes from a number of empirical studies (Skelton 1988b, Adams Smith 1984:28, Gosden 1993:68). Since English is the world's language of scientific communication, much more attention has understandably been paid to it than to other languages, such as Spanish. Only more recently, due to the increasing interest in the study of rhetorical patterns across languages and cultures, can we find some research based on the study of Spanish writing (e.g., see Connor 1996, Valero-Garcés 1996, Moreno 1997, and Burgess 2002). However, to my knowledge, no investigators have yet looked specifically at issues related to creating truly comparable corpora for comparing Spanish and English biomedical writing.

The aim of this paper is to describe how a Spanish biomedical research article corpus was modeled on the one Salager-Meyer used in 1994. The corpus will first be used to carry out a study of Spanish hedging that will be parallel to Salager-Meyer's. Later, attempts will be made to perform other studies of Spanish for comparison to previous research on English biomedical discourse. I will describe why it was not possible to duplicate certain aspects of Salager-Meyer's corpus, and therefore, why I had to change certain criteria. I will argue that the resulting corpus will nevertheless provide a valid basis for research comparing the features of Spanish biomedical research papers to the features that have been described for English language articles, the starting point for which will be a replication of the study Salager-Meyer carried out in 1994. Then, after relating it with other comparison studies in the field, my Spanish corpus could also be used for further study of other rhetorical features in medical discourse and to identify possible differences across languages and genres.

1.1 The concept of hedging

> "Hedges are words or phrases whose job is to make things fuzzier."
> (Lakoff 1972)

The earliest definition of the term was given by Lakoff in 1972 and since then many other authors have been studying this rhetorical phenomenon in different languages and genres (Salager-Meyer 1994; Hyland 2000, and Piqué *et al.* 2002 among others). However, after reviewing the literature available on the topic we have observed that most of the studies have been focused on the English language and the way scientific writers modulate their discourse in this language. A possible explanation of this fact could be that nowadays there is a dominance of English in scientific research publications, hence the need of non natives to publish in English to maintain visibility within the discourse community. We conclude this section with another definition by Salager Meyer of this rhetorical feature: By resorting to such expressions, researchers can avoid absolutes and thus indicate exactly the degrees of certainty with which they present their conclusions and also how strongly they want to align themselves with their claims. Room for disagreement is provided in this way.

1.2 Hedging functions in scientific discourse

As we have already seen in the previous section of this paper, hedging could be defined as modulating one's writing by moderating and softening the exactness of a factual statement. Expressions of this type have as their aim either:

1) to allow scientists to present their knowledge cautiously and introduce claims,
2) to encourage dialogue with the audience and facilitate discussion,
3) to qualify categorical commitment or
4) to avoid author's involvement due to the impossibility to reach absolute accuracy of facts.

2. Methodology

To achieve our goal a corpus of 20 different RAs was selected from six different Spanish medical journals: 10 original RPs (Research Papers) and 10 CRs (Case Reports).

2.1 Criteria

First, I found that unlike the high impact factor journals in English Salager-Meyer had selected for her study (*Annals of Internal Medicine, Archives of Internal Medicine, The British Medical Journal, The Lancet and The New England Journal of Medicine*), there were no such high impact factor journals available in Spanish. Second, I thought that in order to have a corpus large enough to reveal major trends, it was better to select 20 samples of biomedical research articles, rather than the 15 Salager-Meyer had previously used. Third, being the RP one of the most relevant features in the rhetoric of medical discourse, it seemed more than appropriate to me to have a larger number of samples from this major genre (10 RP instead of 5 RP). And finally, considering the general interest in the modern state of the language, the articles selected for the Spanish corpus

TABLE 1. Comparison of two corpora of biomedical research articles in two languages (Spanish and English)

OLIVER DEL OLMO (2003)	*SALAGER-MEYER (1994)*
RAs IN SPANISH	**RAs IN ENGLISH**
20 BIOMEDICAL RAs FROM 6 NON HIGH IMPACT FACTOR JOURNALS	**15 BIOMEDICAL RAs FROM 5 HIGH IMPACT FACTOR JOURNALS**
10 RP AND 10 CR ARTICLES PUBLISHED BETWEEN 1999 AND 2002	**5 RP AND 10 CR ARTICLES PUBLISHED BETWEEN 1980 AND 1990**

were published between 1999-2002, whereas Salager-Meyer's belonged to the period between 1980-1990. All the above mentioned changes in the selection criteria for modelling a valid Spanish corpus are summarized in the following table.

2.2 *Steps*

After the above mentioned changes in the selection criteria, I took the following steps in creating a Spanish corpus of biomedical research articles:

Step 1.1 asked specialists on the field of Spanish medical publishing (doctors, editors and translators) to identify the best journals published in Spanish in the field of biomedicine. And as a result, I obtained the following selection:

-Archivos de Bronconeumologia
-Atención Primaria
-Medicina Clínica
-Revista Española de Anestesiología y Reanimación
-Revista Española de Cardiología
-Revista de Neurología

Step 2. I contacted some of the editors of the above selected journals in Spanish, through an e-mail letter (see Figure 1 below), where I asked them to name 6 RAs (3 RP and 3 CR), which they considered to be examples of "good writing" in their most recent published issues (the value of the articles being the high quality of their writing not their content). In that letter the author of this paper introduced herself, stated the aim of such compilation of research articles in biomedicine (to carry out a linguistic analysis of specific rhetorical features in two different medical genres) and indicated the specific characteristics those articles should contain: 1) types of genre (RP and CR), 2) good quality of writing, and 3) approximate article's period of publication (2000-2002).

FIGURE 1. E-mail letter to the Spanish biomedical journal editors

Apreciado Sr. Director:

Me llamo Sonia Oliver del Olmo y soy profesora de inglés para fines específicos en la Universidad Internacional de Cataluña y en la Politécnica. Como ya le habrá comentado [nombre del contacto], en estos momentos estoy recopilando información para mi tesis doctoral y me remito a usted para que me recomiende tres artículos científicos y tres casos clínicos que a su parecer sean buenos y que se hayan publicado en su revista en los últimos dos años.

Este trabajo de investigación para la Universidad Pompeu Fabra (Facultad de Traducción y Filología) se centrará en el análisis lingüístico de algunas de las características más relevantes de la retórica del discurso

científico de carácter médico en lengua castellana y por ello, estoy muy interesada en poder utilizar como corpus para esta tesis una muestra de buenos ejemplares de este tipo de género textual.

Le estaría muy agradecida si usted pudiera sugerirme dichos ejemplares, enviármelos por *e-mail* o indicarme la manera más fácil de acceder a ellos. Es posible que dadas las características del estudio fuera muy útil tener estos documentos en formato electrónico. Y no hace falta mencionar que esta información será tratada con la máxima confidencialidad y respeto al anonimato de los autores.

De antemano agradecerle su atención e interés en el proyecto. No dude en contactar conmigo si surgiera cualquier duda o cuestión sobre el tema.

Muy cordialmente,

Sonia Oliver del Olmo

Step 3. Out of a total of 30 RAs (RP and CR) recommended by some of the editors of the six different journals in Spanish, 20 articles were selected at random so as to obtain a balance number among the different journals. See Table 2.

Step 4. Each article was codified according to the medical genre it belonged to (RP or CR) and it was labelled with SP (for Spanish language). See Tables 3 and 4.

Step 5. Finally, the corpus was digitised for further research. In this procedure the following equipment was required*, a scanner (HP Scanjet 6300 Q, Sofware OCR (Optical Character Recognition) -Caere Omnipage Pro 10.0, True Page (multiple column configuration) and the compression tool: Windzip 8.1, being the operative system Windows 2000 Professional and the Text treatment Word 1997/2000.

3. Results

In Table 2,1 have included the classification of the RAs according to: a) the Spanish scientific journal in which they were published, b) the language of the journal (as I might include an extra corpus of NNs (non-native speakers) biomedical research articles in English in a further study), c) the number of articles per journal included in the Spanish corpus, d) the genre of the article selected, and e) the article's year of publication.

In Table 3 and Table 4,1 present the Spanish biomedical research articles classified according to the medical genre they belong to (Table 3 for RP and Table 4 for CR). Both tables include the same features: a) the article's numerical code, b) the scientific journal where each article was published, c) the Institution where the investigation was carried out, and d) the article's total number of running words.

TABLE 2. Spanish corpus of RAs in biomedicine

SCIENTIFIC JOURNAL (Abbreviation)	LANGUAGE	NUMBER OF ARTICLESIN THE CORPUS	ARTICLE* SGENRE: RESEARCHFAPER7 CASE REPORT	YEAR
Arch. Bronconeumol	Spanish	3	1 Research Paper 2 Case Reports	1999 1999/2000
Aten Primaria	Spanish	2	Research Papers	2001
Med Clin (Bare)	Spanish	5	2 Research Papers 3 Case Reports	2000 2000
Rev Esp Anestesiol reanim	Spanish	5 .	2 Research Papers 3 Case Reports	2000/2001 2000/2001
Rev Esp Cardiol	Spanish	2	Research Papers	2002
Rev Neurol	Spanish	3	1 Research Paper 2 Case Reports	2001 2001

4. Discussion

While creating a valid corpus of Spanish biomedical research articles, I could observe some similarities and some differences with the one used by Salager-Meyer in 1994 to study hedging in English medical discourse.

As for the **similarities**, I could see that both corpora:

1) Belong to the field of biomedicine.
2) Study two different kinds of genre (RP and CR).
3) Consist of full-length research articles.
4) Are (specifically) created to study hedging devices.

But I also found out some **differences** as a result of the changes in the selection criteria for the Spanish corpus and those differences can be summarized as follow:

1)Spanish journals do not have high impact factors (unlike Salager-Meyer's selection of journals for the English corpus).
2)The Spanish corpus consists of 10 RP and 10 CR, whereas Salager-Meyer's consisted of 5 RP andlO CR.
3) The Spanish articles were selected among six recommended Spanish journals in biomedicine while Salager-Meyer only used five journals in English to create her corpus of biomedical research articles.
4) The Spanish articles were published between 1999-2002 and Salager-Meyer's were published in the period between 1980-1990.

TABLE 3. Coding of the Spanish Research Papers(RP) in biomedicine

CODE	JOURNAL	INSTITUTION WHERE RESEARCH WAS CARRIED OUT	RUNNING WORDS
RP-SP1	*Revista Española de Anestesiología y Reanimación.*	Hospital Mar-Esperanfa. IMAS. Barcelona. Hospital Universitari Dr. Josep Trueta. Girona	4,997 words
RP-SP2	*Revista Española de Anestesiología y Reanimación.*	Hospital Universidad Virgen de la Arrixaca. Murcia. Facultad de Medicina. Murcia.	4,149 words
RP-SP3	*Atención Primaria.*	Centro de Salud de Estella. Servicio Navarro de Salud Osasunbidea.	4,075 words
RP-SP4	*Atención Primaria.*	Centro de Salud La Solana. Talavera de la Reina. Toledo.	2,430 words
RP-SP5	*Medicina Clínica.*	Hospital Clínic. Barcelona.	6,342 words
RP-SP6	*Medicina Clínica.*	Hospital Universitario Miguel Servet. Centro de Salud Actur Sur. Zaragoza.	7,004 words
RP-SP7	*Revista Española de Cardiología.*	Hospital Dr. Negrín de Gran Canaria. Las Palmas.	6,206 words
RP-SP8	*Revista Española de Cardiología.*	Hospital Universitario La Fe. Valencia.	7,324 words
RP-SP9	*Archivos de Bronconeumología.*	Hospital Universitario de la Princesa de Madrid. Clínica Ruber. Madrid.	5,825 words
RP-SP10	*Revista de Neurología*	Hospital La Paz. Universidad Autónoma de Madrid.	5,268 words

TABLE 4. Coding of the Spanish Case Reports (CR) in biomedicine

CODE	JOURNAL	INSTITUTION WHERE RESEARCH WAS CARRIED OUT	RUNNING WORDS
CR-SP1	*Medicina Clínica.*	Hospital de la Santa Creu i Sant Pau. Barcelona	2,753 words
CR-SP2	*Medicina Clínica.*	Hospital Valí d'Hebron. Barcelona. Facultad de Medicina. Murcia.	3,319 words
CR-SP3	*Medicina Clínica.*	Hospital de Terrasa. Barcelona. Fundación Jiménez Díaz. Madrid.	4,259 words
CR-SP4	*Revista de Neurología.*	Hospital Sant Joan de Déu-Clínic. Hospital Universitario Sant Joan de Déu. Espluges de Llobregat, Barcelona.	2,108 words
CR-SP5	*Revista de Neurología*	Hospital Clínico Universitario San Carlos. Madrid.	4,933 words
CR-SP6	*Archivos de Bronco-neumología*	Hospital La Paz. Universidad Autónoma de Madrid.	1,518 words
CR-SP7	*Archivos de Bronco-nuemología*	Hospital Clínic. Barcelona.	1,916 words
CR-SP8	*Revista Española de Anesteseología y Reanimación.*	Hospital General Universidad La Paz. Madrid.	3,241 words
CR-SP9	*Revista Española de Anesteseología y Reanimación.*	Hospital General Universitario de Valencia.	2,151 words
CR-SP10	*Revista Española de Anesteseología y Reanimación.*	Hospital General Universitario José M. Morales Meseguer. Murcia.	3,342 words

5) The Spanish corpus was digitised (as specified in Step 5 in the methodology section of this paper).
6) Spanish (not English) is the language of my corpus of biomedical research articles.

Sinclair claimed in 1996 that a corpus was "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of language " and that was my aim in creating a valid corpus of Spanish biomedical research articles. By developing a Spanish corpus consisting of full-length articles recommended by specialists and published in respected Spanish medical journals in the field, I could, somehow, guarantee the quality of the samples collected and thus, make the findings of my study more linguistically comparable to those with established impact factors in English language publications. First, the creation of this Spanish corpus will allow me to replicate the study Salager-Meyer carried out in 1994, before going on to other comparative studies. Then, my Spanish corpus could also be used for further study of other rhetorical features in medical discourse and to identify possible differences across languages and genres.

ACKNOWLEDGEMENTS

For helpful comments on this corpus project, I thank those who attended its presentation at CILFE6, especially John. M. Swales, Tim Johns, Laurence Anthony, Jordi Piqué and Santiago Posteguillo. I would also like to thank Carmen López Ferrero and Mary Ellen Kerans for advice on designing the corpus. Erika Ehnis, Eva Cid, Claudia Barahona, Lasse Olsen and many others gave their support and advice.

References

HYLAND, K. (1994). *Hedging in Academic Writing and EAP Textbooks.* English for Specific Purposes Vol 13, N° 3, pp. 239-256.

HYLAND, K. (2000). *Disciplinary Discourses. Social Interactions in Academic Writing.* London: Longman.

LAKOFF, G. (1972) Hedges: A Study in Meaning Criteria and the Logic of Fuzzy concepts. *Journal of Philosophical Logic,* 2, pp. 458-508.

PIQUE, J. et al. (2002). "Epistemic and Deontic Modality: A Linguistic Indicator of Disciplinary Variation in Academic English". *LSP & Professional Communication* 2, 2, pp. 49-65.

SALAGER-MEYER, F. (1994). "Hedges and Textual Communicative Function in Medical English Written Discourse", *English for Specific Purposes* 13, 2, pp. 149-70.

SWALES, J. (1990). Genre Analysis. Cambridge, UK: Cambridge University Press.

Appendix A

Corpus of biomedical research papers (RP) in Spanish (the number is used for referential purposes as indicated in Step 4 in the Methodology section of this paper)

RP-SP1 Santiveri X, Castillo J, Navarro M, Pardina B, Villalonga A, Castaño, J. "Remifentanilo o propofol para la sedación en anestesia subaracnoidea. Efectos sobre la ventilación, estabilidad hemodinámica e índice biespectral". *Rev Esp Anestesiol Reanim* 2001 ;48, pp. 409-414.

RP-SP2 Hernández-Palazón J, Tortosa Serrano, J.A., García-Palenciano C, Molero Molero E, Burguillos López S, Pérez Flores D. "Respuesta cardiovascular a la intubación traqueal en pacientes con tumor intracraneal. Estudio comparativo entre el urapidilo y la lidocaina". *Rev Esp Anestesiol Reanim* 2000;47, pp. 146-150.

RP-SP3 Dura Travé T, Mauleón Rosquil C, Gúrpide Ayarra N. "Valoración del estado nutricional de una población adolescente (10-14 años) en atención primaria. Estudio evolutivo (1994-2000)". *Aten Primaria* 2001:28(9), pp. 590-594.

RP-SP4 Labrador García M.S, Merino Segovia R, Jiménez Domínguez C, García Salvador, Segura Fragoso A, y Hernández Lanchas C. "Prevalencia de Fibrilación auricular en mayores de 65 años de una zona de salud". *Aten Primaria* 2001:28(10), pp. 648-651.

RP-SP5 Ortega M, Esteban María J, Miró O, Sánchez M y Millá J. "Estudio prospectivo de los enfermos que abandonan un servicio de urgencias antes de ser atendidos por el médico ". *Med Clin* (Bare) 2000:115, pp. 15-20.

RP-SP6 Gomallón F, Valdepérez J, Garuz R, Fuentes J, Barrera F, Malo J, Tirado M, Simón MA. "Análisis coste-efectividad de dos estrategias de erradicación de Helicobacter pylori: resultados de un estudio prospectivo y aleatorizado en atención Primaria". *Med Clin* (Bare) 2000:115, pp. 1-6.

RP-SP7 Hernández-Ortega E, Medina Fernández-Aceituno A, Rodríguez Esparragón Francisco J, Hernández Perera O, Melián Nuez F, Delgado Espinosa A, Fíuza Pérez D, Anabitarte Prieto A, y Rodríguez Pérez José C. "Relevancia de los polimorfismos génicos del sistema renina-angiotensina en la enfermedad coronaria". *Rev Esp. Cardiol.* 2002;55(2), pp. 92-9.

RP-SP8 Osea J, Quesada A, Amau Miguel A, Osa A, Hervás I, Almenar L, Palencia M, Mateo A y Algarra F. "Péptido Cerebral Natriurético .Valor diagnóstico en la insuficiencia cardiaca". *Rev Esp Cardiol.* 2002;55(1), pp. 7-15.

RP-SP9 Espinosa de los Monteros, M. J., González, A, Rodríguez, F, Gabriel, R y Ancochea, J. "Análisis descriptivo (características clínicas y funcionales) de la población asmática de un área sanitaria". *Arch. Bronconeumol* 1999; 35, pp. 518-524.

RP-SP10 Martínez-Bermejo, A, López-Martín, V, Arcas, J, Tendero, A, Roche, C, Alvarado, F, Ruza, F. "Coma alfa: correlación clínica, electroencefalográfica y etiológica en la edad pedriática". *Rev. Neurol.* 2001; 33 (12), pp. 1101-1105.

Appendix B

Corpus of Biomedicai Case Reports (CR) in Spanish

CR-SP1 Corominas H, Domènech M, González-Juan D, González-Suárez B, Díaz C, Pujo J, Vázquez G y Baiget M. "Aplasia medular tras administración de azatioprina: papel del polimorfismo genético de la tiopurina metiltransferasa *Med Clin* (Barc.)2000;115, pp. 299-301.

CR-SP2 Martínez-Valle F, Anton Capdevila J, Ribera E, Pigrau C, y Pahissa A. "Seudo tumor pancreático de etiología tuberculosa". *Med Clin* (Bare.) 2000; 115, pp. 460-462.

CR-SP3 Martínez Gimeno M, Trujillo MJ, García Sandoval B, del Rio T, Maseras M,, Ayuso C y Carballo M. "Mutación Asp-190-Tyr en el gen de la rodopsina en una familia española afectada de retinosis pigmentaria autosómica dominante *Med Clin* (Barc.)2000, pp. 115:699-703.

CR-SP4 Serrano M, Campistol J, Chávez B, Caritg J, Fortuny C, Costa J.M. "Tuberculomas intracraneales múltiples en la infancia". *Rev. Neurol.* 2001;33(1), pp. 44-46.

CR-SP5 Santiago R, Domínguez M, Campos-Castelló J. "Infarto cerebral en la infancia corno complicación de migraña con aura. A propòsito de un caso". *Rev. Neurol.* 2001;33(12), pp. 1143-1148.

CR-SP6 Ramírez, T, Prados, C, Gómez de Terreros Caro, J, Villamor, J y Álvarez-Sala, R. "Hemosiderosis pulmonar idiopàtica en paciente de edad avanzada ". *Arch Bronconeumol.* 1999;35, pp. 507-509.

CR-SP7 Torrego Fernández, A, Santos Pérez, S, Brea Folco, J, Barberà Mir, J.A. y Ricardo Vallés, C. "Disfimción de cuerdas vocales simulando asma inducida por el ejercicio". *Arch Bronconeumol.* 2000; 36, pp. 533-535.

CR-SP8 Hemández-Ganzedo, C, Pestaña, D, Burgueño, M. D., de la Quintana, B y Criado, A. "Anestesia en un caso de sindrome de Holt-Oram". *Rev. Esp. Anestesiol. Reanim.* 2001; 48, pp. 434-437.

CR-SP9 Errando. C.L. "Ritmo nodal tras la administración de atropina a pacientes bradicárdicos bajo anestesia subaracnoidea. Descripción de cuatro casos y revisión fisiopatológica y terapéutica". *Rev. Esp. Anestesiol. Reanim.* 2000; 48, pp. 384-386.

CR-SP10 Pérez, J.A, Padilla, J, Rodríguez, M.A., Cura, S, Sánchez, Cutillas, M.J. y Sanz, J.

"Esplenectomía en paciente con betatalasemia intermedia y anemia hemolítica grave ". *Rev. Esp. Anestesiol Reanim.* 2001;48, pp. 288-291.