EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 438–446
CILC2016. 8th International
Conference on Corpus Linguistics

Methodological Considerations on the Design and Compilation of a Spanish learner corpus

An Vande Casteele[1] and Kim Collewaert[1]
[1] Vrije Universiteit Brussel, Brussels, Belgium
An.Vande.Casteele@vub.ac.be, Kim.Collewaert@vub.ac.be

Abstract

As a corpus is a representation of the linguistic reality, it is important to have homogeneous, quantifiable and valid data. This article aims at discussing the issue of elaborating a corpus of oral data from language learners of Spanish. We hereby do not merely focus on the data collection, but also on the difficulties that arise regarding the experimental design, the selection of the participants, the elaboration of a transcription model and the analysis of the data. The discussion will be based upon our own research project, for which oral samples from Spanish language learners of different proficiency levels have been collected in order to be analysed cross-sectionally. Furthermore, this article focuses on the oral experiment specifically designed for this project, similar to those of previous studies on similar subjects. Next to this, we will also discuss the procedure used for the transcription of the data and finally, a codification system will be elaborated.

1 Introduction

The present article on the design of a corpus of oral data of L2 Spanish learners arises out of a PhD-project proposal on the use of referential mechanisms by language learners of Spanish. Several issues regarding the research's method and strategy had to be addressed and finally led to a coherent and fully elaborated project. In this process, defining the methodology of the research proved to be the most challenging part. Consequently, it has been adapted several times, as will be discussed later in this article. As the study's experimental design correlates with the methodology, the conceptualization of the project also experienced some adaptations throughout the elaboration process. Besides these two topics, the present article will also comment on possible difficulties that the data collection, the selection of the participants and the data analysis can pose. In sum, this paper offers a theoretical reflection, describing the methodological process behind the elaboration of an L2 Spanish learners' corpus.

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 438–446

2 The development of a research project based upon a Spanish learner corpus

As we aim at discussing the development of our research project on the use of referential mechanisms by language learners of Spanish, first an overview of the main objectives of our study will be provided. Then, the participants will be presented and finally, the most challenging components, namely the study's design and methodology, will be dealt with.

2.1 What? Aim and objectives of the study

One of the first steps in the elaboration of a research project is the delineation of the study's purpose and objectives. From the onset, it was clear that the study would focus on the acquisition of Spanish as a second language, in accordance with earlier research on the use of discourse markers by language learners (Vande Casteele & Collewaert 2013). More specifically, the goal of the research project is to investigate the use of the linguistic mechanisms for referring to persons in the oral L2 Spanish discourse of native speakers of Dutch, and this, by means of a discourse-structural study of a self-compiled corpus. The study will be particularly relevant, as previous investigations that have examined aspects of reference in L2 Spanish by Dutch native speakers are still thin on the ground, apart from De Haan *et al.* (2008, 2006) who paid attention to the use of cohesive devices such as connectors and reference words in a study on the development of Dutch speakers' writing competence in L2 English and L2 Spanish. Moreover, previous studies on the use and development of reference in L2 Spanish have mainly focused on the use of specific and separate kinds of referring expressions in relative isolation, such as, for instance, the use of pronouns (Malovrh & Lee 2013, Jegerski *et al.* 2011, Rothman 2009). In contrast, our study will present a comprehensive and integrated approach, addressing the broad repertoire of referential devices that L2 learners have at their disposal and investigating entire referential chains rather than separate referential expressions.

As to reference, it is generally assumed that the choice of a referential expression is primarily determined by its salience or prominence in a discourse. As such, the different kinds of referring expressions are selected depending on their assumed degree of accessibility, i.e. the extent to which a referent is (supposed to be) available and identifiable to the addressee at any particular stage of the discourse. As Arnold & Griffin (2007: 522) state, speakers thus "select reference forms based upon their assumptions about the addressees' mental states".

In particular, our study has three main objectives. First, it aims at investigating to what extent the learners' choice of referring expressions correspond to the predictions of the accessibility scales proposed by Ariel (1990) – for English – and (2002) – for Spanish. These scales, shown below, represent the repertoire of referring expressions as a continuum from the most (e.g. zero anaphora) to the least accessible expression (e.g. a descriptive noun phrase).

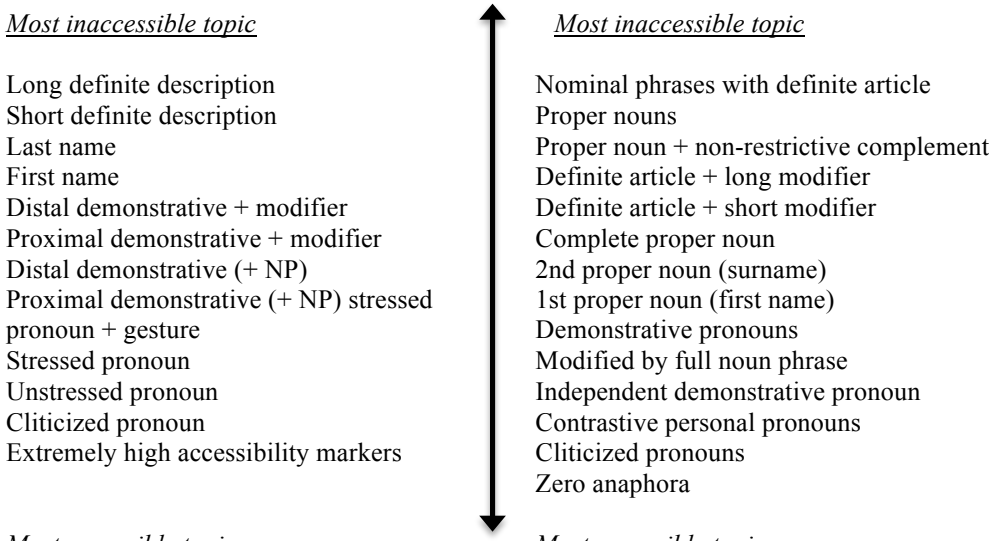

Figure 1: Accessibility marking scale (Ariel 1990)

Figure 2: Accessibility marking scale (Figueras 2002)

As such, we want to verify if the learners first use lowly accessible markers before turning to highly accessible markers. Therefore, the entire referential chains will be analysed, which will enable us to investigate possible correlations between the use of low accessible markers and the introduction of a referent on the one hand, and between the choice of high accessible markers and referent maintenance on the other hand. After identifying all the different referring mechanisms, their overall distribution in the discourse will be examined in terms of the parameters that previous studies have proposed as influencing referential choice, for example the referent's topicality in the discourse, the distance to previous references or the presence of intervening referents (Arnold & Griffin 2007; Van Vliet 2008; Vázquez Rosas 2004; Vande Casteele 2014). The results of the language learners of Spanish will be compared with those of a control group of Spanish native speakers.

Second, the study intends to investigate the overspecification phenomenon, i.e. the use of expressions that "contain more information than necessary for the unique identification of the referent" (Arts *et al.* 2011: 362). In this context, a recent pilot study (Vande Casteele & Collewaert 2016) indicated that the L2 Spanish oral discourse of Dutch-speaking learners contained a significant number of expressions potentially indicative of this language feature, for example the unnecessarily and abundantly repetition of the proper name within a same sequence, instead of the use of more accessible expressions, such as the zero anaphora. As an illustration, examples (1) and (2):

(1) "Y Beatriz se ha desmayado/están en la oficina/y una mujer habla con Beatriz/y otra vas a traer agua y una bebida con alcohol para ayudar Beatriz/ Álvaro y Diego entran y la mujer dice que Beatriz no se sienta bueno" (NN201410)

(2)"Hay una discusión entre Diego y Álvaro y pega a Diego pienso pero de

nuevo Beatriz es en lugar de Diego entonces pega a Beatriz y intenta de cuidar a Beatriz" (NN201402)

Moreover, overspecification constitutes an important notion in studies on reference in both L1 and L2 perspectives (Arts *et al.* 2011; Koolen 2013; Leclercq & Lenart 2013; Van Vliet 2008, 2002).

Third, the study will take into account that reference is considered an "addressee-oriented process" (Koolen 2013: 27), as could be observed in the definition of accessibility mentioned above. As such, the role of the addressee within the experimental setting could possibly influence the participants' referential choices. Consequently, we will experimentally investigate this hypothesis for L2 discourse by means of two different experimental settings in which the contrasting variable is the addressee, i.e. one setting in which the referent can be assumed to be 'known' to the addressee, and one setting where this assumption does not hold.

2.2 Who? Research participants

As previously mentioned, the research participants in our study are Dutch-speaking language learners of Spanish. They will be selected from among the students of the Linguistics & Literature and Applied Linguistics Departments at the Vrije Universiteit Brussel. The selection will be based on a language background profile, which will give us insights on the participants' knowledge of other languages than Spanish and on the participants' exposure to Spanish outside the formal classroom setting, as these aspects may influence or explain individual differences in the overall L2 proficiency development and in the use of referring mechanisms in their L2.

Eventually, oral production data from four groups of students will be collected. The four groups coincide with the curriculum (that is three Bachelor years and one Master year). The L2 Spanish participants will also take a standardized general Spanish L2 proficiency test as an independent measure of their overall Spanish proficiency development. Finally, the same retell data will also be collected from 50 native Spanish speakers to serve as a control group for interpreting and analysing the results from the L2 learners.

2.3 How? Research design and methodology

The development of the methodological section of the research project has proved to be the most challenging part, and as it is closely related to the study's design, several changes have been discussed. First, we will present the task that has been designed in order to be able to test the use of referential mechanisms in the oral discourse of language learners of Spanish. While previous studies on similar subjects used story-telling methods such as the Frog Story (cf. Yusun Kang 2004) or Charlie Chaplin's Modern Times (*cf.* Jarvis 2002), this study will forego these more traditional forms of story-telling by following Watorek (2004) and Leclercq & Lenart (2013) who used a story-telling task based on a film fragment. However, our task will be slightly different, since it will not be based on a silent cartoon, but on a compilation of extracts from the Spanish telenovela *Yo soy Bea*, in which appear various characters. In our study, the consecutive references to each of them will be investigated. The main characters in the compilation are Álvaro, the new handsome manager of a fashion journal, and Beatriz, his smart, but ugly secretary. All other characters appearing in the fragments are linked to Álvaro: his best friend Gonzalo, his brother-in-law and enemy Diego, his pretty girlfriend Cayetana and his parents Francisco and María. A recently conducted pilot study has shown that this particular procedure and task yield rich and relatively unmonitored and unplanned data for investigating referent introduction and maintenance on the one hand, and reference shifts on the other hand. Moreover, as the video contains seven fragments, the narratives can also be divided into seven sequences, which will allow us to analyse the consecutive references within these

sequences. We assume thus that each part responds to the structure presented in the accessibility marking scales (Ariel 1990; Figueras 2002), as a new sequence should imply the start of a new referring chain.

With respect to the task procedure, the assignment will be administered to each participant individually and takes about 20 minutes to complete. It consists of a first viewing session of the video compilation including audio (in Spanish), permitting the participants to have a general understanding of the storyline. Next, the participants are asked to retell the story in their own words to a researcher-interviewer while watching the compilation a second time, this time without the sound track. This procedure allows the story telling time to run quasi-parallel to the video viewing time, to prevent that participants will only produce a brief summary of the story rather than actually retell it (as has often been the case in previous studies using picture stories or video retellings), thus yielding richer data sets. Other advantages of this technique include the impossibility of participants to monitor or edit their oral discourse after the task (as is often the case with written narrative data) and the unavailability of aid of external resources (e.g. dictionaries, thesauri, reference grammars), as a researcher is present to monitor the task. Finally, another characteristic is the time pressure on the participants, as they have to retell the story while keeping pace with the unfolding actions on the video (again limiting their opportunities to monitor their utterances).

As we already mentioned, the study's design has undergone various changes throughout the elaboration process, as we had to choose between a cross-sectional or a longitudinal design. In an initial design, the participants would have been divided into two proficiency cohorts: an early language learners' cohort with students of the first Bachelor and an advanced learners' cohort with students of the third Bachelor. Both students' groups would have been tested twice: after being tested in their first and third Bachelor year, respectively, the two cohorts would repeat the task when they were in their second Bachelor year and in their Master. As such, this research design would have allowed us to compare the data of the two proficiency cohorts cross-sectionally and, moreover, to investigate the data of each cohort longitudinally, but we had to be aware of the possibility of a certain participant attrition, which could reduce the number of participants during the second test moment.

Next, a slightly different configuration was designed in order to extend the interval between the two test moments, as to ensure that sufficient gains in L2 proficiency could be expected. As such, the first cohort would have been tested in their first and third Bachelor year, and the second cohort in their second and Master year. Again, the results could have been analysed both cross-sectionally and longitudinally, but the possibility of participant attrition only increased. This procedure would require for the researchers more years to obtain the data and consequently would considerably delay the analysis.

Thirdly, the study could be designed exclusively longitudinally. Longitudinal studies provide data about the same participants at different points in time and are, consequently, useful to observe development or change in the target group's results. As such, a longitudinal study is able to show patterns of a certain variable over time and to expose cause-and-effect-relationships. In this setting, a group of Dutch-speaking first year students then would have been tested five times throughout their three Bachelor years. As we decided to also take into account the experiment setting as a possible variable for referential choice, the test would have taken place in two different settings and consequently, our group of participants would have been divided into two groups, who would each perform the task in a different setting. Nevertheless, in the case of our study, such longitudinal design would have posed several difficulties regarding the data and the data collection. First, it would have been necessary to use multiple video fragments in order to test the students and to avoid habituation to the test procedure. Moreover, we could not assure that the same characters would appear in each fragment and if so, that they would appear with the same frequency. Naturally, this could compromise the comparability and hence the statistical validity of the data.

Consequently, we ultimately decided to opt for a, traditionally called, cross-sectional design, or what Granger (2016) suggests to call a "quasi-longitudinal study", i.e. "a study that contains data which are gathered at a single point in time, but from learners with identified proficiency levels". Moreover, Granger (2016) argues that this type of study can draw interesting insights on language acquisition. As we intend to use a standardized language proficiency test as an independent measure of the overall Spanish proficiency development of our participants, our study could thus be considered "quasi-longitudinal". This design allows us to obtain a homogeneous set of data as the same video fragment can be used throughout the entire study, which will result in comparable and statistically valid data. It also increases the feasibility of the project, as it allows us to test a higher number of participants. It is less subject to participant attrition and offers results in a shorter time-lapse. In sum, four groups of language learners (coinciding with the curriculum, i.e. three Bachelor years and one Master year) will eventually be tested and their results will be compared. Next to this, the results of the language learners will also be compared with those of a control group of Spanish native speakers.

As already mentioned, the addressee plays a significant role in the referential process, which can be considered as "addressee-oriented" (Koolen 2013: 27). Consequently, the study aims at implementing this key feature and will verify to what extent the role of the addressee within the experiment setting influences the participants' referential choices. Therefore, each of the four tested groups will be divided in two. The two groups will perform the same retell task but under two different experimentally manipulated conditions, in which the addressee assumes two different roles and statuses. In the first condition, the participant and the research-interviewer (the addressee) will first both watch the video clip during the first viewing session (with the soundtrack) together. Then the participant will be asked to retell the story (while watching the clip a second time, without sound) to the 'knowledgeable' interviewer (who is familiar with the story line and the characters). In the second condition, the participants will be invited to retell the story to a "naïve interlocutor" (cf. Leclercq & Lenart 2013) who was not present during the first viewing and who pretends to be unfamiliar with the story and its characters. This design will allow us to investigate to which extent the mental state of the addressee influences the participants' choice of referring expressions. In addition to the L2 Spanish data, the Spanish native speakers will also be divided into two groups. So, the first group will retell the story to a knowledgeable interlocutor (condition 1) and the second one to a naïve interlocutor (condition 2).

Throughout the data collection, the oral productions of the participants will be audio-recorded and then transcribed, annotated and coded in CHAT format, allowing for semi-automated analysis with the CLAN programmes (MacWhinney 2000). As an illustration, two examples of transcriptions that have been realized during the pilot study: example (1) forms part of the learners' corpus, example (2) belongs to the native speakers' corpus.

(1) *Ah y aquí hay un conflicto entre los dos hombres pero no he entendido el la razón y euhm*
Álvaro euhm entento euhm intenta a euhm golpear el otro hombre pero euhm
Es Beatriz que es irrida- irridado pero euhm puedes ver que el hombre es guapo pero es también feliz es euhm
(silencio largo)
Pero no he entendido mucho que de se que lo dice
(silencio largo)
La novia es eh no le gusta que es tan feliz
Es euhm un poco como su hermano euhm
(silencio largo)
Y aquí los dos hermanos euhm (silencio) charlan pero no saben que Beatriz les escuchan y euhm también se trato de un complotto eh creo
Euhm una mentira
Y creo que se trato también de Beatriz de su puesto como secretaria secretaría
No no le gusta que se qué qué los hombres dicen euhm y

(silencio largo)
No creo
Se trata de (silencio) el hombre se euhm (silencio largo) euhm (silencio) sem- semblar de estar enamorada por esto como esto
euhm pero no he entendido mucho
es euh
(silencio largo)
Es euhm más el hermano que le puso a hacer cosas que Álvaro normalmente no hace euhm (NN141506)

(2) *Durante las primeras imágenes presentan a los personajes*
y Beatriz ésta escribiendo un correo
en donde explica que a ella le gustaría haber sido ser ministro de economía
pero que se conforma con por lo menos ser secretaria
una fea secretaria
después sale la escena donde Álvaro y Gonzalo están dando la bienvenida a las modelos
y Álvaro se comporta de manera muy cariñosa con cada una de las modelos
las las mira fijamen las mira detalladamente y en un momento se encuentra con Beatriz
en donde se comporta sorprendido por pues por su apariencia
después llaman en llaman en la recepción a Beatriz y a otra de las muchachas para la entrevista para secretarias
y la persona que la llama le hace el comentario de que la imagen es lo principal en la empresa
ella se queda callada y se siente un poco mal acerca del comentario silencio (N151603)

After the compilation of the corpus, it will be necessary to create some tools to analyse the referential mechanisms in the oral discourse of our participants. As such, we will have to develop an appropriate annotation system for our data, as well as to reflect on which statistical procedures will be used to analyse the data. First, the references to persons in the transcribed Spanish interlanguage data will be analysed and coded by means of an annotation scheme that will be developed on the basis of the coding scheme designed for our pilot study and advanced in other proposals in the literature, such as the annotation scheme for German proposed in Bauman and Riester (2010, 2012, 2013) and in Riester et al. (2010), and the coding system by Andreou et al. (2015) for referring expressions in Greek-German children's data. This coding system analysed the morpho-syntactic form of the expressions (e.g. (in)definite noun phrase, null pronoun, object clitics, personal pronoun, etc.) on the one hand, and the discourse function (i.e. if the referring expression is used for character introduction, maintenance or reintroduction), on the other. Importantly, our coding system will capture formal (morpho-syntactic and lexical) as well as semantic-pragmatic and discourse-functional aspects of reference to persons in our database. In general, the basic units of analysis will be entire referential chains, rather than isolated referential expressions. Second, in order to answer the study's research questions, the coded data will be analysed both qualitatively and quantitatively, the latter by means of appropriate statistical procedures. Given the categorical nature of the envisaged data, including a binary outcome variable (low vs. high accessibility), Pearson's chi-squared test, rank correlation test and logistic regression modelling are likely to be suitable methods of data analysis.

3 Final considerations

The aim of this paper was to present some reflections on the elaboration of a research project on the use of referential mechanisms in the oral discourse of language learners of Spanish. As such, the

participants would be Dutch-speaking students of L2 Spanish at the Vrije Universiteit Brussel (VUB). The research objectives, then, were quite evident from the onset of the study, also thanks to a pilot study previously conducted on the subject. Nevertheless, the study's design proved to be the most challenging part of the process and different options had to be explored and examined. The combination of a cross-sectional and a longitudinal design as well as an exclusively longitudinal study have been considered, but seemed rather inappropriate to our study, as they imply the use of multiple video fragments, which would comprome the comparability and the validity of the data and the statistical analysis. Furthermore, these designs present a higher risk of participant attrition and require a longer period of time and a consistent follow-up. Consequently, a quasi-longitudinal design has been opted for, since this procedure presents several advantages. First, it allows for the use of the same video fragment throughout the study, so creating comparable and statistically valid data. Second, it benefits the feasibility of the study, as the data collection is less time-consuming and less prone to participant attrition.

References

Andreou, M., Knopp, E., Bongartz, C. & Tsimpli, I. (2015). Character reference in Greek-German bilingual children's narratives. In Roberts, L., McManus, K., Vanek, N. & Trenkic, D. (eds.) EuroSLA Yearbook Volume 15 (pp. 1-40). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ariel, M. (1990). Accessing noun-phrase antecedents. London: Routledge.

Arnold, J. & Griffin, Z. (2007). The Effect of Additional Characters on Choice of Referring Expression: Everyone Competes. Journal of Memory and Language, 56, 521-536.

Arts, A., Maes, A., Noordman, L. & Jansen, C. (2011). Overspecification facilitates object identification. Journal of Pragmatics, 43, 361-374.

Baumann, S. & Riester, A. (2013). Coreference, lexical givenness and prosody in German. Lingua, 136, 16-37 (Special Issue 'Information Structure Triggers', Hartmann, J., Winkler, S. & Radó, J. (eds.)).

Baumann, S. & Riester, A. (2012). Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Elordieta, G. & Prieto, P. (eds.) Prosody and Meaning. Interface Explorations 25, Berlin: Mouton de Gruyter.

Baumann, S. & Riester, A. (2010). Annotating information status in spontaneous speech. In Proceedings of the Fifth International Conference on Speech Prosody, Chicago.

De Haan, P. & Van Esch, K. (2008). Measuring and assessing the development of foreign language writing competence. Porta Linguarom, 9, 7-21.

Figueras, C. (2002). La jerarquía de accesibilidad de las expresiones referenciales en español. Revista española de lingüística, 32, 1, 53-96.

Granger, S. (2016). Learner corpora and foreign language learning and teaching. Symposium within the framework of the Francqui-chair, Kortrijk, 24 March 2016.

Jarvis, Scott (2002). Topic continuity in L2 English article use. Studies in Second Language Acquisition, 24, 387-418.

Jegerski, J., Van Patten, B. & Keating, G.D. (2011). Cross-linguistic variation and the acquisition of pronominal reference in L2 Spanish. Second Language Research, 27, 481-507.

Koolen, R. (2013). Need I say more? On overspecification in definite reference. PhD Dissertation. Tilburg University.

Leclercq, P. & Lenart, E. (2013). Discourse Cohesion and Accessibility of Referents in Oral Narratives: A Comparison of L1 and L2 Acquisition of French and English. Discours, 12.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk (3rd Ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Malovrh, P. & Lee, J. (2013). The Developmental Dimension in Instructed Second Language Learning. The L2 Acquisition of Object Pronouns in Spanish. London: Bloomsburry.

Riester, A., Lorenz, D. & Seemann, N. (2010). A recursive annotation scheme for referential information status. In Proceedings of the Seventh International Conference of Language Resources and Evaluation (LREC) (pp.717-722), Valletta, Malta.

Rothman, J. (2009). Pragmatic deficits with syntactic consequences?: L2 pronominal subjects and the syntax–pragmatics interface. Journal of Pragmatics, 41, 951-973.

Vande Casteele, A. (2014). Referent accessibility of Appositive Constructions in Spanish press. Bulletin of Hispanic Studies, 91.1, 1-18.

Vande Casteele, A. & Collewaert, K. (2013). The use of discourse markers in Spanish Language Learners' Written Compositions. Procedia: Social and Behavioral Sciences, 95, 550-556.

Vande Casteele, A. & Collewaert, K. (2016). The use of referring expressions in Spanish language learners' oral narratives: an exploratory study. In Alonso Almeida, F., Ortega Barrera, I., Quintana Toledo, E. & Sánchez Cuervo M.E. (eds.) Input a Word, Analyze the World: Selected Approaches to Corpus Linguistics (pp. 415-424). Newcastle upon Tyne: Cambridge Scholars Publishing.

Van Esch, K., De Haan, P., Frissen, L., González Santero, I. & De La Torre Miranda, A. (2006). Evolución en la competencia escrita de estudiantes de español como lengua extranjera. Estudios de Lingüística Aplicada, 43, 55-76.

Van Vliet, S. (2008). Proper nouns and pronouns: the production of referential expressions in narrative discourse. PhD Dissertation. Netherlands Graduate School of Linguistics.

Van Vliet, S. (2002). Overspecified NPs marking conceptual shifts in narrative discourse. Linguistics in the Netherlands 2002, 187-198.

Vázquez Rosas, V. (2004). Algunas reflexiones sobre el cálculo de la distancia referencial. D.E.L.T.A., 20(1), 24-47.

Watorek, M. (ed.) (2004). Langages: Construction du discours par des enfants et des apprenants adultes 155. Paris: A. Colin.

Yusun Kang, J. (2004). Telling a coherent story in a Foreign Language. Analysis of Korean EFL learners' referential strategies in oral narrative discourse. Journal of Pragmatics, 36, 2, 1975-1990.