ride. A review journal for digital editions and resources

published by the IDE

Rezension des „Corpus Oral de Referencia de la Lengua Española Contempóranea“

CORLEC, Universidad Autónoma de Madrid (ed.), 1992. http://www.lllf.uam.es/ESP/Corlec.html (Last Accessed: 30.04.2018). Reviewed by Katrin Betz (Universität Bamberg), katrin.betz (at) uni-bamberg.de.

Abstract

In this paper we review the „Corpus Oral de Referencia de la Lengua Española Contempóranea“. This corpus is a carefully compiled text collection of orthographically transcribed recordings of oral conversations. As it was compiled as a reference corpus, it provides texts of different styles and registers and has a considerable size of about 1,100,000 words. CORLEC therefore is an important resource for researchers interested in the field of spoken language. The Corpus is freely available as SGML-based-files or plain-text-files. As both variants are similar in their structure and content, the review addresses the two formats. The review first provides some background information and an overview of the structure of the corpus. Then, the transcription files are described in their structure and content. Eventually, a short overview of the integration of the corpus into other projects and a résumé are given.

Einleitung

1 Das „Corpus Oral de Referencia de la Lengua Española Contempóranea“ (im Folgenden CORLEC genannt) war das erste (digital verfügbare) größere Korpus der

spontanen, gesprochenen Sprache des Spanischen (Moreno-Sandoval 2002, 2005). Zudem ist es bis dato das größte frei verfügbare Korpus des gesprochenen Spanisch. Das Korpus wurde 1991-1992 von der Universidad Autónoma de Madrid mit Unterstützung des IBM am Lehrstuhl für Allgemeine Sprachwissenschaft unter der Leitung von Francisco A. Marcos Marín erstellt. Erhältlich ist das Korpus über die Internetseite des „Laboratorio de Lingüística Informática“ der Universidad Autonóma unter http://www.lllf.uam.es/ESP/Corlec.html[1], wo es als SGML-basiertes-Format oder als plain-text-Dateien im zip-Archiv zur Verfügung gestellt wird.

Verfügbarkeit

2 Das Korpus kann laut Lizenzangaben frei zu wissenschaftlichen Zwecken weiterverwendet werden und ist hierfür in vollem Umfang zugänglich. Neben der Verfügbarkeit des Korpus als zip-Archiv auf der oben genannten Internetseite scheint es keine weiteren Bezugsquellen für das Korpus zu geben. Unklar ist dabei, ob das Korpus auch langfristig auf der Internetseite angeboten werden wird.

Versionen des Korpus

3 Das Korpus wird mittlerweile in zwei Versionen angeboten: als ASCII kodierte SGML-basierte Daten (Standard Generalized Markup Language) und als UTF-8 kodierte plain-text-Dateien. Bei den SGML-basierten Daten handelt es sich um das ursprüngliche Korpus, die neuere plain-text-Version ist erst in den letzten Jahren zur Verfügung gestellt worden (leider wird für keine der Versionen angegeben, seit wann sie zur Verfügung stehen). Der Aufbau der Dateien dieser beiden Korporaversionen und der Inhalt sind zwar gleich, sie unterscheiden sich aber in der Art der Kodierung von Metadaten, Sprecheinsätzen und sprachlichen Merkmalen. In dem vorliegenden Review werden deshalb bei der Beschreibung des Aufbaus und der Grundstruktur des Korpus sowie der weiteren Auszeichnungen beide Versionen zusammen besprochen. In einem folgenden Punkt wird dann auf die Unterschiede eingegangen.

Hintergrund: ein Referenzkorpus aus den Neunzigern

4 Wie oben erwähnt, entstand das CORLEC Anfang der Neunziger. Seit dieser Zeit ist die technische Entwicklung stark vorangeschritten und auch die Sprachwissenschaft hat sich weiterentwickelt: das an der gesprochenen Sprache entstandene Interesse hat sich weiter verfestigt und etabliert, die Sozio-, die Variations- aber vor allem die

Korpuslinguistik haben zunehmend an Bedeutung gewonnen. Zudem haben sich interdisziplinäre Zweige entwickelt und die Verwendung digitaler Ressourcen und Methoden ist mittlerweile ein fester Bestandteil der linguistischen Methodik (vgl. z.B: Gries 2009, Biber 2015). Deshalb soll hier vorweggenommen werden, dass bestimmte Merkmale und Eigenschaften des CORLEC, die heute als Mangel wahrgenommen werden, zu der Zeit seiner Erstellung wohl dem Standard entsprachen oder aber innovativ waren. Das Korpus ist daher in seiner Zusammensetzung und Größe nur bedingt mit modernen Korpora wie dem „C-Oral-Rom“, dem „Corpus del Español“ von Mark Davies (2002-) oder dem „CREA“ der Real Academia Española (2008-) vergleichbar. Ziel dieses Reviews ist es deshalb, das CORLEC im Hinblick auf seine Struktur, seine Daten und seine möglichen Verwendungen zu beschreiben, ohne dabei einen wertenden Vergleich mit moderneren Korpora anzustreben.

Basisdaten des CORLEC

5 Das CORLEC wurde als Referenzkorpus der gesprochenen Sprache konzipiert. Als solches soll es die gesprochene Sprache als Ganzes repräsentieren und Texte aus möglichst allen relevanten diasystematischen Varietäten abbilden. Dabei sollte das Korpus ausgeglichen sein und verschiedene Varietäten in einem sinnvollen Verhältnis enthalten (EAGLE 1996). Gerade bei Korpora der gesprochenen Sprache erweist sich die Erstellung eines ausgeglichenen und balancierten Korpus aber als schwierig. Einerseits sind die Begriffe ‚ausgeglichen‘ und ‚balanciert‘ eher vage:

> The notion of balance is even more vague than representativeness, but the word is frequently used, and clearly for many people it is meaningful and useful. Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements.
>
> (Sinclair 2004)

6 Andererseits sind, abgesehen von der Vagheit der Begriffe, solche Verteilungen im Bereich der gesprochenen Sprache nur schwer umsetzbar, da es z.B. wesentlich einfacher ist, Mitschnitte aus einer Radiosendung, einer Fernsehsendung oder einem Vortrag aufzuzeichnen als Aufnahmen aus einem privaten Gespräch. Ebenso ist es leichter und schneller, gehobene Standardsprache zu transkribieren als ein informelles, dialektal gefärbtes Gespräch mit sich überlappendem Turn-Taking. Zusätzlich ist auch zu

berücksichtigen, dass das Korpus mit den technologischen Gegebenheiten der Neunziger erstellt wurde: Für die Erstellung des Korpus wurden mit Hilfe analoger Aufnahmegeräte Aufzeichnungen von mehreren Mitarbeitern des Projektes erstellt. Dabei handelt es sich nicht um zum Zwecke der Aufnahme erstellte Interviews, sondern um authentische Gespräche. Die Gespräche stammen aus unterschiedlichen Quellen. Z.B. wurden private Konversationen aufgenommen und transkribiert, aber auch Konferenzen oder Radio- und Fernsehbeiträge. Berücksichtigt man diesen Hintergrund, wurde ein Korpus beachtlicher Größe und Variation zusammengestellt: Wie oben erwähnt umfasst das Korpus ca. 1.100.000 Wörter, die auf 505 transkribierte Gesprächsdateien verteilt sind.

Grundlegende Strukturen des Korpus

7 Die Gesprächsdateien sind im downloadbaren zip-Archiv auf verschiedene Ordner verteilt. Diese Ordner entsprechen einer thematischen Kategorisierung der Texte. Die thematischen Bereiche, nach denen die Texte sortiert wurden, sind die Folgenden: Administrativos y políticos (Gespräche aus dem Bereich der Verwaltung oder mit politischem Bezug) 5.6%, Científicos (Gespräche mit wissenschaftlichem Bezug) 3.3%, Conversacionales o familiares (Private oder familiäre Gespräche, informellere Gespräche) 24.5 %, Educativos (Gespräche aus dem Bildungsumfeld) 5.3%, Humanísticos (Gespräche aus dem geisteswissenschaftlichen Bereich) 5.6%, Instrucciones (megafonía) (Anweisungen über Lautsprecher/Megafon) 0.6%, Jurídicos (Juristischer Bereich) 3.2%, Lúdicos (Gespräche im Zusammenhang mit Spielen, Wettbewerben, Verlosungen etc.) 5.6%, Periodísticos (Journalistische Gespräche): Debates (Diskussionsrunden) 8.5%, Deportes (Sport) 5.3%, Documentales (Dokumentationen) 2.6%, Entrevistas (Interviews) 15.6%, Noticiario (Nachrichten) 6.6%, Publicitarios (Werbungen) 2.8%, Religiosos (Religiöser Bezug) 1.1%, Técnicos (Technischer Bezug 3.9%) (vgl. Marín o.J.).

```
#archivo: ACON006D
#fecha de grabación: 23-5-91
#fuente: conversación entre amigos, en la calle
#localización: Madrid
#H1: varón, informático, 25 años
#H2: varón, informático, 28 años
#H3: varón, informático, 26 años
#hablante alemán: mujer, economista, 24 años
#términos: Costa rica, vídeos, inventos, tabaco, bebidas, ranas, serpientes
H1: ¿Qué pasa, tío? ¿Qué tal tu sarampión?
H2: Bien... aquí sigue.
H1: Sí, ¿no ibas a ir...?
H3: Maña[palabra cortada]... ¿mañana?
H2: Mañana me la quitan.
```

Abb. 1: Neuere Version (plain text).

```
<cinta 006>
<ACON006D.ASC>
<23-5-91>
<fuente=conversación entre amigos, en la calle>
<localización: Madrid>
<H1=varón, informático, 25 años>
<H2=varón, informático, 28 años>
<H3=varón, informático, 26 años>
<hablante alemán=mujer, mujer, economista, 24 años>
<términos= Costa rica, vídeos, inventos, tabaco, bebidas, ranas, serpientes>
<texto>
<H1>  ¿Qué pasa, tío? ¿Qué tal tu sarampión?
<H2> Bien... aquí sigue.
<H1> Sí, ¿no ibas a ir...?
<H3> Maña<palabra cortada>... ¿mañana?
…</texto>
```

Abb. 2: Ältere Version (SGML-basiert).

8 Bei den transkribierten Texten selbst handelt es sich um eine orthografische Transkription, in der die Interpunktion ergänzt wurde.[2] Die beiden abgebildeten Beispiele (Abb. 1 und 2) aus den beiden Korpusversionen sollen einen kurzen Einblick in die Transkription sowie die Grundstruktur der Daten geben.

9 Die Dateien beginnen jeweils mit den Metadaten. Diese geben grundlegende Informationen zu der Gesprächssituation bzw. zu der Aufnahme: Aufnahmenummer, Aufnahmedatum, Aufnahmeort (z.B. Madrid), Quelle der Gespräche (z.B. Telefongespräch in einer Firma). Außerdem sind auch soziolinguistische Metadaten zu den Sprechern vorhanden: Ungefähres Alter, Geschlecht, Beruf u.Ä. Zusätzlich werden einige Schlüsselwörter kodiert, die den Gesprächsinhalt betreffen.

10 Nach den Metadaten folgt der transkribierte Text, in dem die einzelnen Sprecheinsätze der Sprecher kodiert sind. Die Sprecher sind dabei nach dem Muster H1, H2, H3, etc. benannt. Für manche ‚besonderen' Sprechertypen werden eigene Tagnamen verwendet, wie z.B. encuestado1, encuestado2.

11 Innerhalb der Texte sind zusätzlich noch einige Merkmale kodiert, die die Eigenschaften der gesprochenen Sprache bzw. die Kommunikationssituation betreffen.

Hierzu gehören etwa Informationen, die phatische Merkmale erfassen (z.B. kodiert als *duda*, *llorando*, *sorpresa*, *asombro*), oder Informationen, die das Turn-Taking betreffen. Daneben wurden auch Hintergrund- oder Nebengeräusche u.Ä. kodiert (*música*, *risa*).

12 Die Kodierung kann dabei recht uneinheitlich ausfallen: Hintergrundlachen kann zum Beispiel kodiert sein als [risa] oder [risas], nicht transkribierter Text als [texto no transcrito], [encuesta no transcrita], [texto no transcrito por ininteligibleteléfono].

13 Laut Angaben orientieren sich die Kodierungsrichtlinien der Gespräche zwar teilweise an dem Datenmodell der TEI (vgl. Marín o.J.), jedoch könnte hier eine Anpassung an jüngere Fassungen und auch konsequentere Anpassungen vorgenommen werden. In diesem Zusammenhang ist auch die Erstellung einer XML-Version des Korpus mit zugehörigem Schema und eventuell auch Schematron eine weitere Verbesserungsmöglichkeit des Datenmodells, die die Weiterverarbeitung erleichtern würde.

Metadaten und variationslinguistische Informationen

14 Das Korpus bietet über die eben beschriebene Korpusstruktur sowie über die oben beschriebenen Metadaten zwar sozio- bzw. variationslinguistische Informationen zu den Texten. Diese Informationen sind jedoch hinsichtlich zweier Punkte problematisch: Einerseits ist zu beachten, dass es sich bei der oben beschriebenen, thematischen Einteilung nur um eine mehr oder weniger stark thematische Einteilung handelt. Durch das Kriterium *conversación o familiares* wird beispielsweise die Gesprächsform sowie das Register der Texte erfasst, nicht aber die Thematik der Gespräche. Die als *periodísticos* zusammengefassten Unterpunkte (denen jeweils ein eigener thematischer Ordner entspricht) beschreiben zum Teil eine Gesprächsstruktur (*debates*, *entrevistas*) oder aber Themen (*deportes*). Innerhalb der Ordner können die Gespräche zudem zusätzlich hinsichtlich des Mediums (face-to-face, Radio, Fernseher), der Formalität, der Spontaneität und der Dialogizität (Vorträge, spontane Privatgespräche) variieren (vgl. hierzu auch die Parameter des Modells Mündlichkeit/ Schriftlichkeit von Koch/Österreicher 1985). Z.B. befindet sich in der Kategorie ‚Wissenschaft' ein Gespräch unter Freunden über Dinosaurier und deren mögliche Rekonstruktion über DNA, ein Vortrag für Abiturienten im Planetarium und ein universitärer Konferenzbeitrag. In dem Ordner über Religion befinden sich ein Mitschnitt einer Predigt, eine Radiodiskussion, eine Fernsehdiskussion und ein direktes Gespräch unter Freunden. Analog dazu sind in den Metadaten zwar Informationen über die

Gesprächsquelle vorhanden. Diese liegen aber sehr unstrukturiert vor und erfassen zum Teil verschiedenartige Informationen unterschiedlich detailliert, wie Code 1 zeigt.

```
CIE/ACIE032A: <fuente=charla en el Planetario de Madrid>
HUM/AHUM033B: <fuente=conferencia en el Instituto de la Mujer>
CON/ACON006A: <fuente=casa particular, conversación familiar>
JUR/PJUR005F: <fuente=radio>
```

Code 1: Informationen über die Gesprächsquellen in den Metadaten.

15 Hier wäre es z.B. bei CON/ACON006A sinnvoll, das Attribut ‚fuente' in mehrere einzelne Attribute zu zerlegen: ein Attribut, das den Ort des Gespräches erfasst und ein zweites, das die Gesprächsart und das Medium (z.B. face-to-face/Radio/Fernseher, direkter Dialog/Monolog) erfasst. Zusätzlich könnten hier auch weitere, oben bereits erwähnte Informationen bezüglich der Kommunikationsbedingungen (Spontaneität, Vertrautheit der Gesprächspartner, Öffentlichkeit-Privat) oder Angaben hinsichtlich der diaphasischen Varietät hinzugefügt werden. Daneben ist für manche Analysen auch die geografische Herkunft des Sprechers von Interesse, die im Allgemeinen unklar bleibt.

16 Wünschenswert bei der Auszeichnung sowohl der bereits vorhandenen Informationen als auch zusätzlicher Informationen wäre vor allem aber eine stärkere Vereinheitlichung und Strukturierung.

Neue versus alte Version

17 Wie an den obigen Beispielen ersichtlich ist, wurde für die neuere Version des Korpus nicht nur die Kodierung geändert, sondern es wurden auch die SGML-basierten Tags umgewandelt. Leider konnte auf der Downloadseite des Korpus keine Übersicht oder Beschreibung gefunden werden, die diese oder weitere Änderungen von einer zur anderen Version dokumentiert. Hier werden deshalb diejenigen Änderungen beschrieben, die bisher festgestellt werden konnten, ohne dabei Vollständigkeit zu gewährleisten. Der auffälligste Unterschied zwischen den Korpora ist die oben erwähnte Umformulierung der SGML-basierten Tags. Für die neue UTF-8-Version wurden die Tags nach folgenden Regeln umformuliert:

- Metadaten-Tags beginnen mit ‚#' und enden mit ‚:' (`<fuente>` wird zu: `#fuente:`)
- Tags zur Sprecherauszeichnung enden mit Doppelpunkt: <H1> wird zu `H1:`

- Leere inline-Elemente werden in eckigen Klammern erfasst: `<risas>` wird zu `[risas]`
- Bei leeren inline-Elementen zur Angabe phatischer Merkmale entfällt der Tag-Name: `<fático=afirmación>` wird zu `[afirmación]`
- inline-Elemente, die Text enthalten, werden von zwei Angaben in eckigen Klammern umrahmt: `<extranjerismo>...</extranjerismo>` wird zu `[extranjerismo]… [fin de extranjerismo]`
- Die Markierung von im Vergleich zur Standardsprache fehlenden Lauten entfällt anscheinend vollständig: `me lo había quita<(d)>o` wird zu `me lo había quitado ese día`.

Integration in andere Korpora

18 Das CORLEC ist nicht nur als Download erhältlich, sondern wurde auch in die beiden größten Online-Korpora der spanischen Sprache integriert: Es ist sowohl im „Corpus del Español“ von Mark Davis als auch im „Corpus de Referencia del Español Oral“ der Real Academia Española enthalten. Es ist dabei nicht klar, welchen Anteil das CORLEC bei diesen Korpora an den jeweiligen Subkorpora der gesprochenen Sprache ausmacht. Ebenso ist es nicht möglich, das Korpus unabhängig von den anderen Teilkorpora abzufragen. Diese Überschneidung sollte bei Arbeiten, die das „Corpus de Referencia del Español Oral“ der RAE und das „Corpus del Español“ einbeziehen wollen, berücksichtigt werden.

Gesamteindruck

19 Das Korpus unterscheidet sich durch seine Größe und seine freie Verfügbarkeit bereits positiv in wichtigen Punkten von anderen Korpora: Das „C-Oral-Rom‘, das „Corpus del Español‘ oder auch das „Corpus de Referencia del Español Oral‘ sind nur auf begrenzte Art und Weisen verfügbar und nutzbar. Das „Valesco‘ Korpus ist zwar online frei erhältlich, verfügt aber nur über 120.246 Wörter (On Mié 2013).

20 Ob das CORLEC letztendlich zur Bearbeitung einer bestimmten Forschungsfrage passend und ergiebig ist, ist stark von den wissenschaftlichen Interessensgebieten und der Methodik abhängig. Bei qualitativen Arbeiten, für die mit einfachen Suchanfragen gezielt nach Sprachbelegen gesucht werden kann, oder allgemein für Arbeiten, die

keinen maschinellen Zugriff auf die Daten erfordern, ist es u.U. angemessener auf eines der online-Korpora zurückzugreifen.

21 Für Arbeiten, die Zugriff auf die gesamten Daten benötigen und nur frei verfügbare Korpora verwenden können/wollen, ist das CORLEC eine passende (und oft auch die einzige) Möglichkeit (für neuere Arbeiten auf der Basis des CORLEC vgl. beispielsweise Betz 2016, Garcia-Marchena 2015, Hennemann 2015). Beide Versionen des Korpus können mit überschaubarem Aufwand in maschinenlesbare XML-Dateien oder menschenlesbare Formate transformiert werden und sind damit sowohl maschinell als auch manuell auswertbar. Die ursprünglichen Audioaufnahmen sind allerdings nicht erhältlich.

22 Abschließend sei erwähnt, dass eine Aufbereitung und Modernisierung des Korpus durchaus lohnenswert und wichtig wäre: ein Erhalt des Korpus ist nicht nur für aktuelle synchrone sprachwissenschaftliche Untersuchungen von Interesse, sondern auch für zukünftige Arbeiten, in denen das Korpus für diachrone Untersuchungen herangezogen werden kann.

Anmerkungen

1. Archivierter Link: https://web.archive.org/web/20171101122930/http://www.lllf.uam.es/ESP/Corlec.html.

2. Für eine ausführliche Beschreibung der Kodierungsrichtlinien vergleiche https://web.archive.org/web/20171101122930/http://www.lllf.uam.es/ESP/Corlec.html.

Bibliographie

Betz, Katrin. 2016. *Adverbien und Depiktive im Spanischen als radiale Kategorien. Eine korpuslinguistische Untersuchung im Rahmen der Konstruktionsgrammatik.* Bamberg: University of Bamberg Press (=Bamberger Beiträge zur Linguistik, 14).

Biber, Douglas and Randi Reppen, eds. 2015. *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press (=Cambridge Handbooks in Language and Linguistics).

CREA. El Corpus de Referencia del Español Actual (Oral). Real Academia Española. 2008.

https://web.archive.org/web/20171101172232/http://www.rae.es/recursos/banco-de-datos/crea-oral.

Cresti, Emanula and Massimo Moneglia, eds. 2005. *C-Oral-Rom. Integrated Reference Corpora for Spoken Romance Languages.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Davies, Mark. 2002-. *Corpus del Español: 100 Million Words, 1200s-1900s.* (Historical / Genres). https://web.archive.org/web/20171101172056/https://www.corpusdelespanol.org/.

EAGLES, eds. 1996. Preliminary Recommendations on Corpus Typology. Last modified May, 1996. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Garcia-Marchena, Oscar. 2015. *Phrases Averbales et Fragments de l'Espagnol Oral-Étude de corpus. Linguistique.* Université Paris Diderot (Paris 7). Français. <tel-01541007>.

Gries, Stefan Thomas. 2009. *Quantitative Corpus Linguistics with R. A Practical Introduction.* New York: Routledge.

Hennemann, Anja. 2015. „On Self-Repair in Spanish. A Qualitative Analysis Using CORLEC." *Research in Corpus Linguistics*, 3: 19–26. Asociación Española de Lingüística de Corpus (AELINCO).

Marcos, Marín, A. Francisco. o.J. *CORLEC: Corpus Oral de Referencia de la Lengua Española Contempóranea.* Universidad Autónoma de Madrid. https://web.archive.org/web/20171101122930/http://www.lllf.uam.es/ESP/Corlec.html.

Moreno-Sandoval, Antonio. 2002. „La Evolución de los Corpus de Habla Espontánea: la Experiencia del LLI-UAM." In *Actas de las II Jornadas en Tecnologías del Habla*, edited by Ayuso, Antonio Rubio, 00-00. Spain (Granada), Dezember 2002.

Moreno-Sandoval, Antonio, Gillermo de la Madrid, Manuel Alcántara et al. 2005. „The Spanish Corpus." In *C-Oral-Rom. Integrated Reference Corpora for Spoken Romance Languages*, edited by Cresti, Emanula and Massimo Moneglia, 135 – 161. Amsterdam / Philadelphia: John Benjamins Publishing Company.

On Mié, Adrian. 31.07.2013. *Valesco. 2.0.* https://web.archive.org/web/20180428085553/http://www.valesco.es/?q=noticias.

Sinclair, John. 2005. „Corpus and Text - Basic Principles.“ In *Developing Linguistic Corpora: a Guide to Good Practice*, edited by Martin Wynne, 1-16. Oxford: Oxbow Books. https://web.archive.org/web/20171101171958/http://ota.ox.ac.uk/documents/creating/dlc/.

Factsheet

Resource reviewed	
Title	CORLEC
Editors	Universidad Autónoma de Madrid
URI	http://www.lllf.uam.es/ESP/Corlec.html
Publication Date	1992
Date of last access	30.04.2018

Reviewer	
Surname	Betz
First Name	Katrin
Organization	Universität Bamberg
Place	Bamberg
Email	katrin.betz (at) uni-bamberg.de

General Information		
Bibliographic description	Can the text collection be identified in terms similar to traditional bibliographic descriptions (title, responsible editors, institution, date(s) of publication, identifier/address)? (cf. Catalogue 1.1)	yes
Contributors	Are the contributors (editors, institutions, associates) of the project documented? (cf. Catalogue 1.3)	yes
Contacts	Is contact information given? (cf. Catalogue 1.4)	yes
Aims		
Documentation	Is there a description of the aims and contents of the text collection? (cf. Catalogue 2.1)	yes
Purpose	What is the purpose of the text collection? (cf. Catalogue 2.2)	Research
Kind of research	What kind of research does the collection allow to conduct primarily? (cf. Catalogue 3.1.8)	Qualitative research

Self-classification	How does the text collection classify itself (e.g. in its title or documentation)? (cf. Catalogue 2.3)	Corpus
Field of research	To which field(s) of research does the text collection contribute? (cf. Catalogue 2.2)	Linguistics
Content		
Era	What era(s) do the texts belong to? (cf. Catalogue 2.5)	Contemporary
Language	What languages are the texts in? (cf. Catalogue 2.5)	Spanish
Types of text	What kind of texts are in the collection? (cf. Catalogue 2.5)	Speech transcripts
Additional information	What kind of information is published in addition to the texts? (cf. Catalogue 2.5)	Introduction
Composition		
Documentation	Are the principles and decisions regarding the design of the text collection, its composition and the selection of texts documented? (cf. Catalogue 3.1.1-3.1.3)	yes
Selection	What selection criteria have been chosen for the text collection? (cf. Catalogue 3.1)	Linguistic characteristics
Size		
Texts/records	How large is the text collection in number of texts/ records? (cf. Catalogue 3.1.4)	> 100
Tokens	How large is the text collection in number of tokens? (cf. Catalogue 3.1.4)	> 1 Mio.
Structure	Does the text collection have identifiable sub-collections or components? (cf. Catalogue 3.1.5)	yes
Data acquisition and integration		
Text recording	Does the text collection record or transcribe the textual data for the first time? (cf. Catalogue 3.1.6)	yes
Text integration	What kind of material has been taken over from other sources? (cf. Catalogue 3.1.6)	none

Quality assurance	Has the quality of the data (transcriptions, metadata, annotations, etc.) been checked? (cf. Catalogue 3.1.7)	no
Typology	Considering aims and methods of the text collection, how would you classify it further? For definitions please consider the help-texts. (cf. Catalogue 3.1.8)	Reference corpus
Data Modelling		
Text treatment	How are the textual sources represented in the digital collection? (cf. Catalogue 3.2.1)	Orthographic transcription
Basic format	In which basic format are the texts encoded? (cf. Catalogue 3.2.4)	Plain text
Annotations		
Annotation type	With what information are the texts further enriched? (cf. Catalogue 3.2.2)	Structural information
Annotation integration	How are the annotations linked to the texts themselves? (cf. Catalogue 3.2.2)	Embedded
Metadata		
Metadata type	What kind of metadata are included in the text collection? (cf. Catalogue 3.2.3)	Descriptive
Metadata level	On which level are the metadata included? (cf. Catalogue 3.2.2)	Individual texts
Data schemas and standards		
Schemas	What kind of data/metadata/annotation schemas are used for the text collection? (cf. Catalogue 3.2.4)	none
Standards	Which standards for text encoding, metadata and annotation are used in the text collection? (cf. Catalogue 3.2.4)	other: TEI-oriented (1992)
Provision		
Accessability of the basic data	Is the textual data accessible in a source format (e.g. XML, TXT)? (cf. Catalogue 4.1)	yes
Download	Can the entire raw data of the project be downloaded (as a whole)? (cf. Catalogue 4.2)	yes

Technical interfaces	Are there technical interfaces which allow the reuse of the data of the text collection in other contexts? (cf. Catalogue 4.2)	none
Analytical data	Besides the textual data, does the project provide analytical data (e.g. statistics) to download or harvest? (cf. Catalogue 4.3)	no
Reuse	Can you use the data with other tools useful for this kind of content? (cf. Catalogue 4.4)	no
User Interface		
Interface provision	Does the text collection have a dedicated user interface designed for the collection at hand in which the texts of the collection are represented and/or in which the data is analyzable? (cf. Catalogue 5.1)	no
Preservation		
Documentation	Does the text collection provide sufficient documentation about the project in general as well as about the aims, contents and methods of the text collection? (cf. Catalogue 6.1)	no
Open Access	Is the text collection Open Access? (cf. Catalogue 6.2)	yes
Rights		
Declared	Are the rights to (re)use the content declared? (cf. Catalogue 6.2)	yes
License	Under what license are the contents released? (cf. Catalogue 6.2)	other: scientific usage
Persistent identification and addressing	Are there persistent identifiers and an addressing system for the text collection and/or parts/objects of it and which mechanism is used to that end? (cf. Catalogue 6.3)	none
Citation	Does the text collection supply citation guidelines? (cf. Catalogue 6.3)	no
Archiving of the data	Does the documentation include information about the long term sustainability of the basic data (archiving of the data)? (cf. Catalogue 6.4)	no
Institutional curation	Does the project provide information about institutional support for the curation and sustainability of the project? (cf. Catalogue 6.4)	no

Completion	Is the text collection completed? (cf. Catalogue 6.4)	yes
Personnel		
Editors	Universidad Autónoma de Madrid Francisco, Marcos Marín	
Contributors	Almudena Ballester, Carrillo Carmen, Santamaría García Elena, Pertierra Torreño Otilia, BrandÆo Cardoso dos Santos Pedro, Luis Díez Orzas	